CIPRIAN RĂULEA

STATISTICĂ PSIHOLOGICĂ ŞI PRELUCRAREA INFORMATIZATĂ A DATELOR

CURS INTRODUCTIV
PENTRU STUDENȚII SPECIALIZĂRIILOR
PSIHOLOGIE ŞI ŞTIINȚELE EDUCAȚIEI

2010

TEME PENTRU STUDIU

Cuvânt înainte

Capitolul 1. Evoluția statisticii şi obiectul ei de studiu
- 1.1. Evoluția istorică a statisticii
- 1.2. Obiectul de studiu şi rolul statisticii
- 1.3. Programe-software utilizate în statistica socială şi psihologică
- 1.4. Noțiuni introductive privind utilizarea programului SPSS

Capitolul 2. Noțiuni fundamentale folosite în statistică
- 2.1. Colectivitatea şi unitatea statistică.
- 2.2. Variabile statistice.
- 2.3. Cuantificarea şi măsurarea fenomenelor psihosociale.
- 2.4. Scale de măsură.
- 2.5. Definirea variabilelor statistice cu ajutorul SPSS.

Capitolul 3. Ordonarea, gruparea şi prezentarea datelor statistice
- 3.1. Serii (distribuții) statistice
- 3.2. Gruparea (sistematizarea) datelor
- 3.3. Prezentarea datelor sub formă de tabele
- 3.4. Reprezentarea grafică a datelor statistice
- 3.5. Utilizarea SPSS pentru ordonarea şi gruparea datelor statistice
- 3.6. Utilizarea SPSS pentru prezentarea datelor statistice sub formă de tabele
- 3.7. Utilizarea SPSS pentru reprezentarea grafică a datelor statistice

Capitolul 4. Indicatori ai tendinței centrale
- 4.1. Mediile
- 4.2. Quantilele: mediana, quartilele, decilele şi centilele
- 4.3. Modul
- 4.4. Relația dintre medie, mediană şi modul
- 4.5. Reprezentări de tip *Boxplots*
- 4.6. Utilizarea SPSS pentru calcularea şi reprezentarea indicatorilor de poziție

Capitolul 5. Indicatori ai variației şi indicatori ai formei
- 5.1. Indicatori simpli (elementari) ai variației
- 5.2. Indicatori sintetici ai variației
- 5.3. Indicatori ai formei distribuției
- 5.4. Utilizarea SPSS pentru calcularea indicatorilor variației şi ai formei

Capitolul 6. Distribuțiile statistice
6.1. Distribuția normală
6.2. Distribuții simetrice şi asimetrice
6.3. Distribuții unimodale şi bimodale
6.4. Valori normate (scoruri z)
6.5. Distribuția normală standardizată

Capitolul 7. Inferența statistică
7.1. Delimitări conceptuale
7.2. Probleme de estimare
7.2.1. Semnificația unei medii.
7.2.2. Semnificația frecvenței
7.3. Testarea ipotezelor
7.4. Testele parametrice t şi z
7.4.1. Testele t şi z pentru un eşantion.
7.4.2. Testele t şi z pentru două eşantioane independente
7.4.3. Testele t şi z pentru două eşantioane dependente
7.5. Utilizarea SPSS pentru aplicarea testului t

Capitolul 8. Corelație şi regresie
8.1. Noțiunea de covarianță
8.2. Coeficienții de corelație
8.2.1. Clasificarea coeficienților de corelație.
8.2.2. Formula coeficientului de corelație liniară simplă (Bravais-Pearson)
8.2.3. Reprezentarea grafică a corelației. Liniaritatea relației.
8.2.4. Interpretarea coeficientului de corelație. Mărimea efectului.
8.3. Coeficienți de corelație parametrici
8.3.1. Coeficientul de corelație Pearson r.
8.3.2. Coeficientul r_{bis}
8.4. Coeficienți de corelație neparametrici: coeficientul de corelație a rangurilor Spearman ρ
8.5. Regresia simplă liniară
8.6. Utilizarea SPSS pentru determinarea coeficienților de corelație

Bibliografie

1.

EVOLUȚIA STATISTICII ŞI OBIECTUL EI DE STUDIU

1.1. Evoluția istorică a statisticii
1.2. Rolul şi scopul statisticii
1.3. Programe-software utilizate în statistica socială şi psihologică
1.4. Noțiuni introductive privind utilizarea programului SPSS

1.1. EVOLUȚIA ISTORICĂ A STATISTICII

Pe măsură ce omenirea a evoluat, statistica s-a îndepărtat radical de statutul de „ramură a matematicii aplicate", în zilele noastre, fiind considerată atât o ştiință, o metodă de cunoaştere a realității socio-economice, cât şi o disciplină de învățământ. Evoluția ei a cunoscut numeroase modificări, precizări, transformări în ceea ce priveşte obiectul ei de studiu dar şi din perspectiva instrumentelor, metodelor sale de cercetare. Ca şi alte ştiințe (matematica, de exemplu) şi această disciplină a parcurs drumul lung şi sinuos de la necesitățile practicii la elaborările teoretice.

Lucrări cu caracter statistic, impuse de nevoile conducerii treburilor publice, apar încă din antichitate. În Egipt, Grecia şi Roma antică erau realizate recensăminte destinate evidențierii resurselor umane şi materiale ale statelor respective. Aceste preocupări însă, au fost considerate naive şi preştiințifice, adevăratul înțeles al statisticii, acela de ştiință, datând doar de la jumătatea secolului al XVII-lea.

Prima analiză statistică, în spirit ştiințific, a unor date culese în prealabil, este datorată lui **John Graunt** (1662) care, pe baza datelor extrase din înştiințările săptămânale cu privire la numărul deceselor înregistrate la Londra, a izbutit să tragă concluzii valabile asupra unor fenomene sociale, precum: natalitatea şi mortalitatea, echilibrul numeric ş.a. Prin aceste preocupări el este considerat „părintele" demografiei.

În Anglia, alături de Graunt, titlul de

John Graunt (1620 - 1674)
comerciant englez, preocupat în timpul liber de „fenomenele demografice" din Londra, publică în 1662 articolul *Natural and Political Observations on the Bills of Mortality*. Ideile sale au fost preluate de Sir William Petty şi de astronomul Edmond Halley şi apoi recunoscute de către Societatea Regală Engleză

„inventator" al statisticii i se atribuie şi lui **William Petty** (1623-1687), care introduce conceptul de *„**aritmetică politică**"* definit ca studiul fenomenelor social-economice *„prin intermediul cifrelor, al măsurilor şi greutăţilor*".

Paralel cu aceste prime preocupări s-a creat, în Germania, un curent de gândire care îşi propunea să descrie situaţia diferitelor state constituite la acea vreme din punct de vedere al populaţiei, bogăţiilor, industriei, comerţului şi finanţelor. Această preocupare se apropie mai mult de sensul etimologic al cuvântului statistică: în limba latină „*status*", are sensul de „*stare*" sau „*stat*". Astfel unii autori atribuie germanului **Gottfried Achenwall** (1719-1772) meritul de a fi întrebuinţat pentru prima dată termenul de statistică, dând întâietate ***şcolii descriptive germane***. Spre deosebire de şcoala engleză a aritmeticii politice, care punea accentul pe colectarea cifrelor şi analiza lor, şcoala descriptivă germană era orientată spre alcătuirea de monografii şi spre compararea calitativă a resurselor statelor.

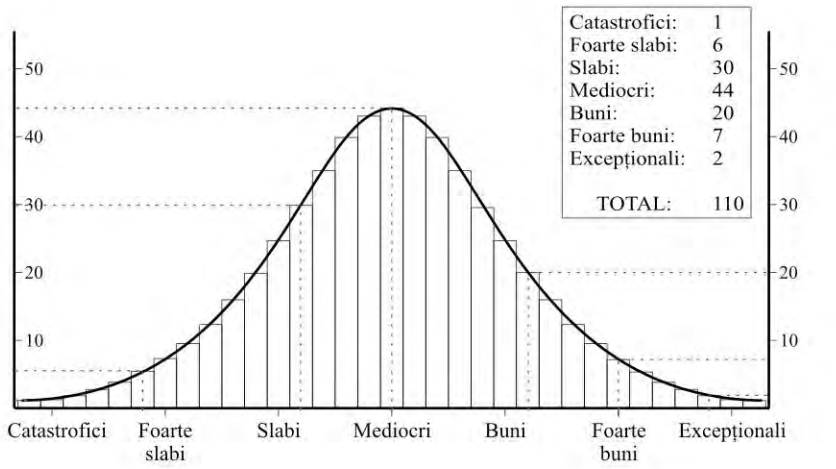
Recunoscând meritul ambelor curente de gândire, T. Rotariu (1999, p.15) consideră că *„ştiinţa statistici, aşa cum arată ea astăzi, datorează aproape totul şcolii engleze, însă contribuţia universitară germană nu poate fi neglijată chiar şi numai pentru motivul că respectivei şcoli îi datorăm numele acestei ştiinţe*".

În spiritul acestei şcoli descriptive, au fost elaborate şi în ţările române în secolele XVIII şi XIX o serie de lucrări ce au contribuit la dezvoltarea statisticii. Prima şi cea mai reprezentativă lucrare de acest gen este „*Descriptio Moldaviae*" (1716) a lui **Dimitrie Cantemir** (1673-1723), o monografie cu caracter geografic, politic, economic, social şi cultural, care îl impune pe autorul ei printre fruntaşii statisticii descriptive europene (D. Porojan, 1993). Şi alţi cronicari precum Grigore Ureche sau Ion Neculce au avut preocupări asemănătoare, iar în 1859, sub domnia lui Alexandru Ioan Cuza, se înfiinţeze primul Birou de Statistică al Ţării Româneşti, condus de Dionisie Pop Marţian (Popescu, 2000)

Karl Friedrich Gauss (1777 - 1855) astronom, matematician şi fizician german. A făcut descoperiri importante în materie de mecanică celestă, electromagnetism, optică. A dezvoltat teoria numerelor. A pus premisele geometriei hiperbolice noneuclidiene

Revenind la începuturile statistici, reamintim faptul că şcoala descriptivă germană era orientată spre descrierea verbală a caracteristicilor statelor, în timp ce aritmetica politică a fost orientată spre analiza fenomenelor sociale şi căutarea legităţilor respective pe baza datelor şi calculelor numerice. Ambele curente au fost depăşite de progresele realizate în domeniul matematicii, în general şi al calculului probabilităţilor, în special. De altfel, dezvoltarea teoriei probabilităţilor a constituit un pas-înainte nu numai pentru statistică, ci şi pentru întreaga creaţie intelectuală a omenirii.

Încă din secolul al XVII-lea s-a observat că măsurătorile repetate ale unui obiect oarecare pot fi reprezentate grafic sub forma unei curbe în formă de clopot. Ecuația curbei normale a fost publicată în 1733 de către **Abraham de Moivre** iar lucrările acestuia au fost dezvoltate ulterior de **Pierre Simon de Laplace** şi **Karl Friedrich Gauss**. În zilele noastre curba normală poartă numele savantului german: clopotul/curba lui Gauss.



Exemplu: Calificativele obținute în urma examenului de statistică de 110 studenți, aleşi aleator.

Odată cu progresele făcute în culegerea datelor şi cu creşterea interesului față de observația şi măsurătorile ştiințifice, statistica a devenit un instrument indispensabil pentru toate ştiințele sociale. Un nume de referință este cel al francezului **Frédéric Le Play** (1806-1870). Acesta este recunoscut prin „introducerea în analiza sociologică a mijloacelor cantitative" (Rotariu *et.al.*, 1999, p.15). Însă, cea mai mare contribuție în această direcție o are belgianul **Adolphe Quételet** (1796-1874), care, la începutul secolului al XIX-lea, aplică teoria probabilităților la studiul fenomenelor sociale, introducând conceptul de „*statistică morală*". Sub inițiativa sa s-a organizat în 1853 primul Congres Internațional de Statistică, la care s-a constituit Institutul Internațional de Statistică.

Adevăratul început al statisticii moderne poate fi fixat la începutul secolului al XX-lea odată cu apariția lucrărilor lui **Karl Pearson** (creatorul statisticii inferențiale sau inductive) şi **Ronald Aylmer Fisher** (a elaborat teoria riguroasă a tragerilor concluziilor din datele observate). Alte nume de referință în fundamentarea statisticii sociale sunt: **C.E. Spearman**, **G.U. Yule**, **M.G. Kendall**, **A.A. Markov**

1.2. OBIECTUL DE STUDIU ŞI ROLUL STATISTICII

În dezvoltarea sa statistica s-a preocupat de acele fenomene şi procese care se produc într-un număr mare de cazuri, denumite **fenomene colective (de masă)** sau, dacă ne referim strict la ştiinţele sociale, **fenomene sociale de masă**. Aceste fenomene de masă se află sub incidenţa legii numerelor mari[1] potrivit căreia variaţiile întâmplătoare de la tendinţa generală se compensează reciproc într-un număr mare de cazuri individuale.

Aplicarea metodelor statisticii în vederea interpretării datelor oferite de observarea fenomenelor de masă permite formularea unor legi statistice. Acestea exprimă media stărilor unei mase de evenimente, tendinţa dominantă care-şi face loc printr-un mare număr de abateri întâmplătoare de la această medie. Legea statistică poate fi evidenţiată numai dacă este supusă observării unui număr suficient de mare de elemente ale ansamblului de studiat (*legea numerelor mari*).

În concluzie, **statistica** studiază aspectele cantitative ale fenomenelor de masă, fenomene care sunt supuse acţiunii legilor statistice şi care se manifestă în condiţii concrete, variabile în timp şi spaţiu.

Încercând o definiţie sintetică, putem afirma că **statistica** reprezintă un **ansamblu de metode şi tehnici utilizate pentru a colecta, a descrie şi a analiza date obţinute în urma unor investigaţii ştiinţifice**.

Statistica a pătruns în toate domeniile ştiinţelor naturii şi ale ştiinţelor sociale, formând discipline de graniţă precum statistica matematică, statistica economică, statistica socială, statistica psihologică, statistica medicală, biostatistica etc. Dintre acestea, aşa-zisa statistică socială şi/sau psihologică se suprapune mult timp şi în mare măsură peste statistica teoretică generală, propunându-şi să culeagă, prelucreze şi să interpreteze informaţiile numerice referitoare la fenomenele psihosociale[2]. Chiar dacă vom folosi de multe ori termenul de statistică socială (sau psihologică), nu considerăm justificată pretenţia unora de a considera statistica socială ca o ştiinţă de sine stătătoare ci, mai degrabă ca o disciplină preocupată de a ilustra modul specific în care statistica generală se aplică în domeniul ştiinţelor sociale şi comportamentale (vezi caseta 1.1.).

Astfel, statistica reprezentând un corp de metode ştiinţifice are rolul de a ne învăţa cum să organizăm observarea fenomenelor de masă şi să obţinem datele necesare, cum să prelucrăm aceste date şi cum să formulăm ipoteze cu privire la relaţiile evidenţiate de aceste date. De asemenea, statistica oferă metode pentru testarea ipotezelor şi pentru confruntarea realităţii cu predicţiile formulate pe baza ipotezelor.

---

[1] Legea numerelor mari a fost formulată de J. Bernoulli în 1713, precizând că într-un număr suficient de mare de cazuri individuale, influenţele factorilor se pot compensa în aşa fel încât să se ajungă la o anumită valoare tipică pentru întreaga colectivitate.

[2] pentru mai mute informaţii vezi Rotariu *et. al.*, 1999, pp. 15-18.

În urma dezvoltării istorice prezentate mai sus statistica modernă s-a separat în două părți distincte dar complementare:

a) ***statistica descriptivă***, se referă la regulile observării statistice directe şi la obținerea informațiilor ce rezultă din prelucrarea datelor empirice. Aici sunt incluse mijloacele clasice ale statisticii: gruparea datelor, distribuțiile de frecvențe, corelația şi regresia, analiza relațiilor dinamice.

b) ***statistica inductivă (inferența statistică),*** se referă la organizarea observării statistice indirecte, prin metode şi tehnici de estimare a însuşirilor unei populații statistice din observații efectuate asupra unei submulțimi de unități statistice, numită eşantion. Include aplicații statistice ale teoriei probabilității.

1.3. PROGRAME-SOFTWARE UTILIZATE ÎN STATISTICA SOCIALĂ ŞI PSIHOLOGICĂ

Cele mai cunoscute programe utilizate de cercetătorii din psihologie, sociologie, asistență socială, economie, pedagogie etc. atunci când realizează analize ştiințifice şi prelucrări statistice complexe sunt: SPSS, SYSTAT, STATISTICA, MINITAB, SuperLab ş.a. Vom descrie pe scurt două din aceste software-uri şi vom prezenta noțiunile de bază necesare utilizării unuia dintre ele (SPSS).

1.4. NOȚIUNI INTRODUCTIVE PRIVIND UTILIZAREA PROGRAMULUI SPSS

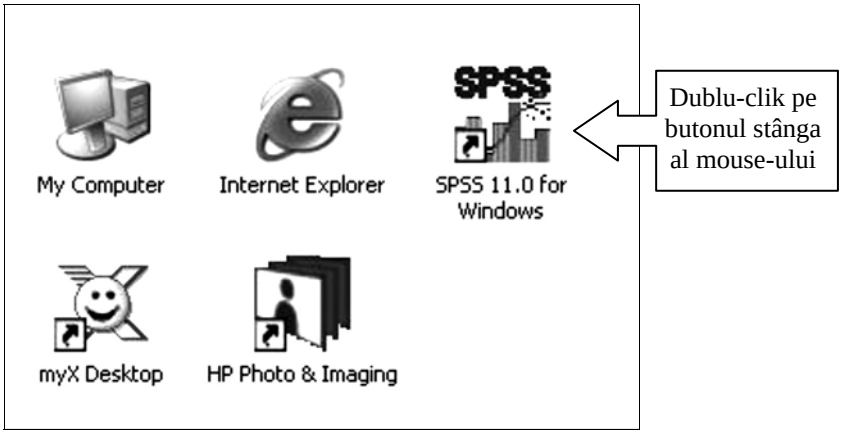
În capitolele aplicative ne vom referi la programul SPSS versiunea 11.0 sub sistemul de operare Windows.[3] Aceste capitole se vor a constitui un ghid de laborator care să-l orienteze şi îndrume pe utilizator în dorința acestuia de a-şi însuşi procedurile şi tehnicile oferite de programul SPSS pentru prelucrarea statistică a datelor.

Deschiderea programului

Pentru pornirea unei sesiuni de lucru în SPSS există următoarele posibilități:

- Daca pe desktop se află shortcut-ul (icon-ul) SPSS se poziționează cursorul pe respectivul icon şi se tastează dublu-clik pe butonul stânga al mouse-ului.

[3] Unele dintre informațiile prezentate nu sunt integrate în versiunile mai vechi (de exemplu, versiunea 7.0) şi sunt diferite sub alte sisteme de operare sau pentru sistemele Macintosh.



- După ce sistemul de operare Windows a fost încărcat, se apasă o singură dată pe butonul stânga al mouse-ului pe următorul traseu:
  Start – Programs – SPSS for Windows – SPSS 11.0 for Windows

După deschiderea programului SPSS, pe ecran va apărea o fereastră de întâmpinare. Este de fapt o fereastră de date (Data View) din cadrul editorului de date (SPSS Data Editor), fără titlu - denumită totuşi „*Untitled*" - şi, atenție!, fără să fie salvată în memoria calculatorului.

- O a treia posibilitate de deschidere a SPSS-ului o reprezintă accesarea (prin dublu-clik) a oricărui fişier acceptat de program.
  Exemple: bazele de date în SPSS sunt fişiere cu extensia *.sav;
  fişierele de tip „syntax" au extensia *.sps;
  fişierele de tip „output" au extensia *.spo etc.

Ferestrele în SPSS

SPSS foloseşte mai multe tipuri de ferestre, fiecăreia dintre ele fiindu-i asociat un anumit tip de fişier. Iată cele mai importante dintre ele:

- **Fereastra de editare a datelor (Date Editor)** se deschide implicit la lansarea unui fişier de tip bază de date, fişier care în SPSS are extensia **.sav*. În această fereastră sunt introduse şi afişate datele de lucru sub forma unui tabel în care liniile reprezintă cazurile (subiecții) iar coloanele variabilele cercetării.

  Fereastra de editare este, la rândul ei, compusă din două foi (ferestre):
  - fereastra de date (**Data View**), folosită pentru introducerea şi vizualizarea seriilor statistice simple (a datelor brute) – vezi figura 1.1.
  - fereastra de gestionare a variabilelor (**Variable View**), folosită pentru definirea şi modificarea variabilelor – vezi figura 1.2.

  Accesarea uneia dintre aceste două ferestre se realizează prin acționarea icon-ului corespunzător din partea stângă-jos a ferestrei de întâmpinare.

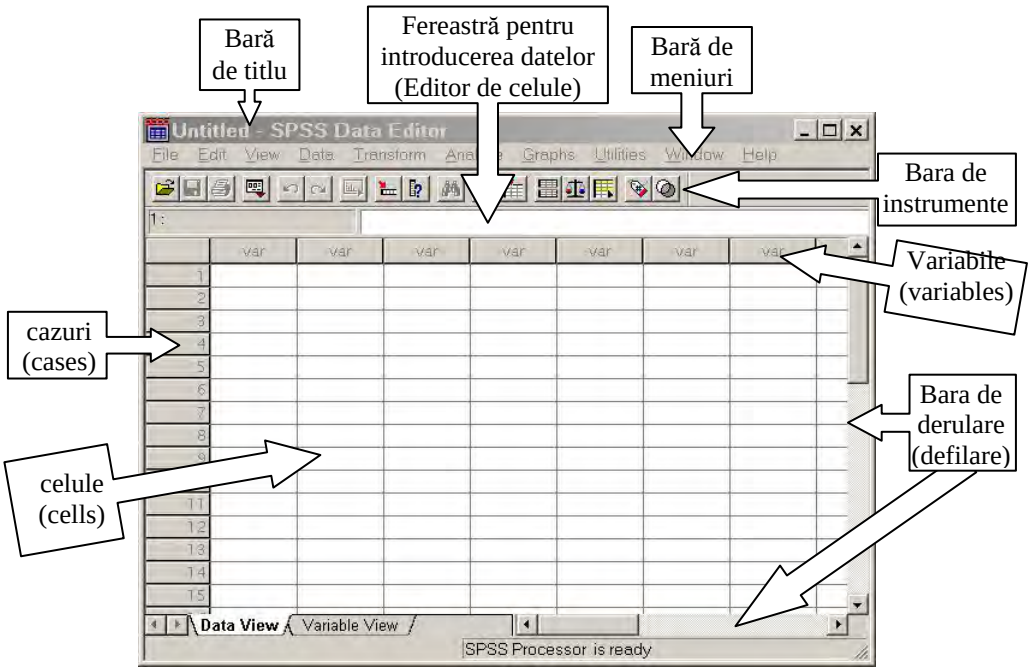


Figura 1.1. *Fereastra de întâmpinare a programului SPSS*

- **Fereastra de gestionare a rezultatelor** sau **Fereastra de ieşire** (**Output – SPSS Viewer**), folosită pentru afişarea şi editarea rezultatelor prelucrărilor statistice (tabele, grafice, indicatori statistici) – vezi figura 1.3. Fereastra Output Viewer este structurată în două cadrane sau zone:
  - cadranul din stânga – *cuprinsul* – prezintă sub forma unei schițe obiectele conținute în fereastră şi
  - cadranul/zona din dreapta – *conținutul* – în care sunt afişate rezultatele obținute prin respectiva analiză.

  Pentru apariția acestei ferestre întâlnim următoarele situații:
  - SPPS deschide automat această fereastră atunci când este solicitat să facă prelucrări şi analize statistice (Atenție: fişierul astfel format va avea denumirea OUTPUTx şi nu este salvat în memoria calculatorului; pentru aceasta trebuie parcurs traseul File - Save sau File - SaveAs);
  - este deschisă de către utilizator prin accesarea unuia dintre fişierele cu extensia **.spo* salvate anterior în memoria calculatorului.

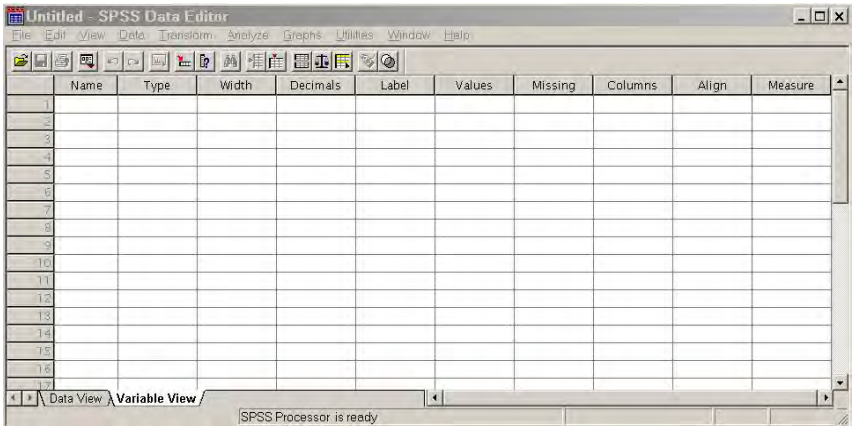

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Name Type Width Decimals Label Values Missing Columns Align Measure
Data View Variable View
SPSS Processor is ready

Figura 1.2. *Fereastra de gestionare a variabilelor*

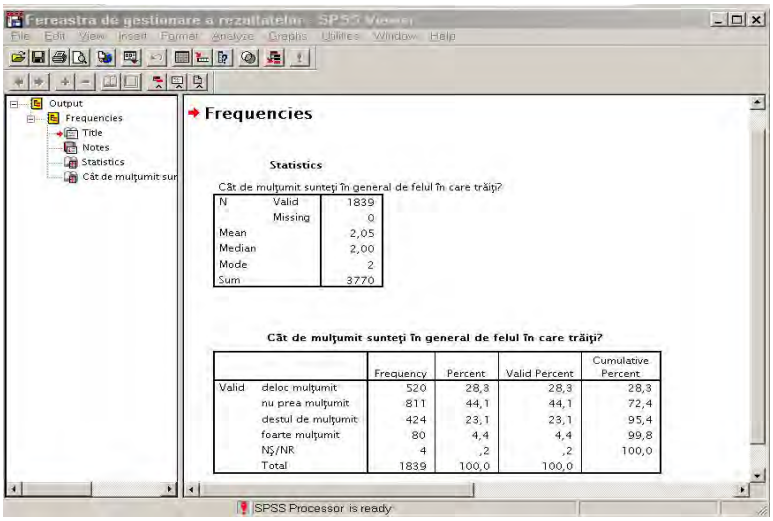

Fereastra de gestionare a rezultatelor - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Output
Frequencies
Title
Notes
Statistics
Cât de mulţumit sun
Frequencies
Statistics
Cât de mulţumit sunteţi în general de felul în care trăiţi?
N Valid 1839
Missing 0
Mean 2,05
Median 2,00
Mode 2
Sum 3770
Cât de mulţumit sunteţi în general de felul în care trăiţi?
Frequency Percent Valid Percent Cumulative Percent
Valid deloc mulţumit 520 28,3 28,3 28,3
nu prea mulţumit 811 44,1 44,1 72,4
destul de mulţumit 424 23,1 23,1 95,4
foarte mulţumit 80 4,4 4,4 99,8
NŞ/NR 4 ,2 ,2 100,0
Total 1839 100,0 100,0
SPSS Processor is ready

Figura 1.3. *Fereastra de gestionare a rezultatelor*

- **Fereastra de editare a comenzilor (Syntax Editor)** permite scrierea comenzilor de către utilizator şi salvarea acestora într-un fişier de tip sintaxă cu extensia **.sps*. Variantele recente ale SPSS conţin meniuri pull-down şi casete de dialog care permit lansarea comenzilor fără a scrie sintaxa acestora.

2.

NOȚIUNI FUNDAMENTALE FOLOSITE ÎN STATISTICĂ

2.1. Colectivitatea şi unitatea statistică.
2.2. Variabile statistice.
2.3. Cunatificarea şi măsurarea fenomenelor psihosociale.
2.4. Scale de măsură.
2.5. Definirea variabilelor statistice cu ajutorul SPSS.

Statistica aplicată în ştiinţele sociale are la bază principiile, tehnicile şi metodele avansate de statistica teoretică generală. Aceasta din urmă, foloseşte un număr mare de noțiuni şi concepte, cu caracter general, care formează vocabularul de bază al statisticii.

În statistica socială, s-au încetățenit de-a lungul timpului, următoarele concepte fundamentale:

- **COLECTIVITATEA (POPULAȚIA) STATISTICĂ** – reprezintă totalitatea elementelor simple sau complexe supuse studiului statistic. (*exemple*: elevii unei şcoli, populația unui oraş)
- **UNITATEA STATISTICĂ (INDIVIDUL STATISTIC)** – reprezintă elementele componente (constitutive) ale colectivităților statistice. Ele pot fi:
  - simple (*exemple*: elevul, studentul, muncitorul);
  - complexe, acestea sunt rezultatul organizării sociale şi economice a colectivității (*exemple*: familia, echipa, clasa de elevi, grupa de studenți).
- **CARACTERISTICA (VARIABILA) STATISTICĂ** – reprezintă însuşirile sau trăsăturile ce definesc şi delimitează unitățile statistice (*exemple*: vârsta, notele şcolare)
- **VALOAREA (VARIANTA),** notată cu **x, y …** – reprezintă forma concretă de manifestare a caracteristicilor la nivelul fiecărei unități statistice (*exemple*: 18 ani, nota 7).
- **FRECVENȚA ABSOLUTĂ,** notată cu f_x, f_y **…** – reprezintă numărul de unități la care se înregistrează aceeaşi variantă (*exemple*: 12 elevii au 18 ani, 3 studenți au obținut nota 7).
- **FRECVENȚA RELATIVĂ (PONDEREA),** notată cu f_{rx}, f_{ry} **…** – se obține prin ponderarea frecvenței absolute, altfel spus, reprezintă procentul unei frecvențe absolute din totalul frecvențelor. (*exemplu*: din 48 de elevii ai unei clase 12 au vârsta de 18 ani, deci ponderea acestora este de 25%)
- **INDICATORII STATISTICI** – reprezintă expresia numerică a unor determinări obiective ce rezultă dintr-o cercetare statistică (*exemple*: media, mediana, abaterea standard).

2.1. COLECTIVITATEA (POPULAȚIA) ŞI UNITATEA STATISTICĂ

După cum am specificat în primul capitol (vezi subcapitolul 1.2.) statistica este preocupată de studierea fenomenelor de masă, a acelor ansambluri finite de elemente care sunt, în mod esențial, de aceeaşi natură calitativă, aparțin aceluiaşi teritoriu şi aceluiaşi timp, altfel spus, sunt statistic omogene. (Jaba & Grama, 2004) Aceste ansambluri sunt cunoscute sub denumirea de *colectivități*, *populații*, *mulțimi*.

COLECTIVITATEA STATISTICĂ (POPULAȚIA STATISTICĂ) – reprezintă totalitatea elementelor simple sau complexe supuse studiului statistic.

În funcție de natura elementelor componente, colectivitățile statistice pot fi formate din ansambluri de ființe, de obiecte sau de evenimente

Exemple: - elevii unei şcoli, populația unui oraş,
- numerele unui anumit ziar apărute într-o lună de zile,
- accidentele rutiere comise pe raza unui județ,
- opiniile electorale înregistrate într-un sondaj.

După numărul elementelor componente, colectivitățile statistice pot fi totale sau parțiale. Primele cuprind totalitatea elementelor componente, în timp ce colectivitățile parțiale, cunoscute sub denumirea de **EŞANTIOANE**, cuprind un număr reprezentativ de unități extrase dintr-o colectivitatea totală. Din acest punct de vedere întâlnim **cercetări exhaustive** - în cazul populațiilor statistice totale - şi **cercetări selective** – ce folosesc proceduri de selecție a indivizilor ce vor incluşi în eşantion.

UNITATEA STATISTICĂ (INDIVIDUL STATISTIC) – reprezintă elementele componente (constitutive) ale colectivităților statistice. Ele pot fi ființe, lucruri, precum şi fapte, evenimente referitoare la acestea.

După gradul de complexitate se clasifică în:

- simple, formate dintr-un singur individ (*exemple*: elevul, angajatul);
- complexe, acestea sunt rezultatul organizării sociale şi economice a colectivității (*exemple*: familia, clasa de elevi, grupa de studenți).

Deşi, atât termenul de individ cât şi cel de populație statistică ne duc cu gândul la natura umană a lucrurilor, exemplele de mai sus pot fi completate cu unități statistice referitoare la lucruri (piesele unui lot supus controlului de calitate) sau la acțiunea omului asupra lucrurilor (măsurarea repetată a unui acelaşi obiect, aruncarea zarului).

2.2. VARIABILE STATISTICE

VARIABILELE STATISTICE (CARACTERISTICILE STATISTICE) – reprezintă însuşirile ce definesc şi delimitează unitățile statistice. Ele exprimă trăsăturile esențiale purtate de unitățile statistice ale unei colectivități, adică dimensiunile prin care se observă, se cuantifică, se măsoară şi înregistrează fiecare unitate din colectivitate. Populațiile umane, cele mai des întâlnite în studiile psihosociale, pot fi caracterizate, de exemplu, prin următoarele variabile: sex, vârstă, nivel de şcolarizare, coeficient de inteligență, tip temperamental ş.a.

Valorile unei variabile statistice se mai numesc **variante** sau **atribute** ale variabilei şi se obţin prin acţiuni concrete de cuantificare şi măsurare a unităţilor unei colectivităţi statistice. De exemplu, variabila „*mediul de proveniență*” are ca variante: *urban* şi *rural*; iar variabila „*notele la examenul de statistică*” are ca valori numerele întregi de la *1* la *10*.

Caracteristicile statistice au proprietatea de a-şi modifica însuşirile în timp şi spaţiu, de la o unitate la alta, în funcţie de influenţele exercitate de o multitudine de factori esenţiali şi întâmplători care acţionează la nivelul fiecărei unităţi din colectivitate. Această proprietate dă variabilelor statistice caracterul de variabilă aleatorie.

În practica de cercetare sunt luate în considerare numai acele variabile care prezintă cel puţin două valori. Dacă, după o anumită caracteristică toate unităţile ar fi identice, aceasta nu ar mai necesita nici un fel de analiză, nemaifiind nevoie să se investigheze cum se manifestă indivizii statistici şi care sunt cauzele acestei variaţii. Să presupunem că toţi studenţii ar obţine nota 10 la disciplina „statistică socială”; nu ar avea nici o relevanţă să verificăm dacă există o legătură între aceste note şi mediile aceloraşi studenţi la examenul de bacalaureat!

Aşadar, cu cât o variabilă îmbracă forme mai diverse, cu atât ea capătă o valoare de cunoaştere mai mare. Numai diversitatea formelor de manifestare a unei însuşiri îi conferă acesteia un interes din partea cercetătorului. (Rotariu *et.al.*, 1999)

- După modul de exprimare, variabilele statistice se clasifică în:
  - **variabile cantitative** (sau **numerice**), exprimate prin numere stabilite prin numărare/măsurare directă sau calcule ulterioare. Numărul stabilit este un *număr cardinal* ce redă intensitatea cu care se manifestă însuşirea respectivă în cazul individului respectiv.
    La rândul lor, variabilele cantitative se clasifică după natura variaţiei în:
    - **variabile discrete**, cu variaţie discontinuă, care pot lua numai valori întregi, de regulă, pozitive. *Exemple*: numărul de membrii din gospodărie, numărul cuvintelor memorate la o probă de memorie.
    - **variabile continue**, cu variaţie continuă, care pot lua orice valoare într-un interval dat. *Exemple*: mediile şcolare anuale, venitul lunar.
  - **variabile calitative** (numite şi variabile **atributive**, **categoriale**, **nominale**), sunt caracteristici ale căror variante de manifestare sunt exprimate atributiv, prin cuvinte. *Exemple*: sexul, mediul de proveniență, tipul temperamental.

Atragem atenţia că într-un studiu statistic sunt reţinute numai acele caracteristici care prezintă interes pentru cercetarea întreprinsă. Pot fi zeci, chiar sute de variabile ce pot caracteriza indivizii unei populaţii statistice. De mult ori ne limităm la a analiza doar câteva dintre ele.

De asemenea, tot cercetătorul este cel care stabileşte, uneori, modul de exprimare şi/sau natura variaţiei unei variabile. O variabilă cantitativă poate fi exprimată calitativ, după cum şi o variabilă cantitativă continuă poate fi transformată, prin rotunjire, într-o variabilă discretă. Exemplul clasic în susţinerea observaţiilor de mai

sus este cel al variabilei „vârstă”: exprimată în ani-luni-zile reprezintă o variabilă cantitativă continuă, exprimată în ani împliniți este o variabilă cantitativă discretă, iar atunci când folosim categoriile tânăr-adult-vârstnic, avem o variabilă calitativă.

În fine, nu trebuie uitat faptul că de foarte multe ori variantele sau atributele variabilelor calitative sunt codificate cu ajutorul numerelor. Aceste coduri reprezintă nişte identificatori, acordarea lor fiind pur convențională, deci ele nu se supun operațiilor matematice sau prelucrărilor statistice bazate pe operații matematice (Jaba & Grama, 2004). De exemplu, întrebarea „Vă place cursul de statistică socială?” poate fi codificată prin 0–NU şi 1–DA sau „Starea civilă” poate fi codificată prin 1-necăsătorit, 2-căsătorit, 3-divorțat, 4-văduv, 5-alte variante; în ambele exemple ar fi inutilă calcularea mediei, a abaterii standard sau a oricărui alt indicator rezultat în urma unor calcule matematice.

2.3. CUANTIFICAREA ŞI MĂSURAREA FENOMENELOR PSIHOSOCIALE

De foarte multe ori în sferă ştiinţelor sociale şi comportamentale rezultatele obținute în urma unor demersuri empirice sunt exprimate calitativ. Partidul cu care a votat un alegător, tipul temperamental al unui manager sau calificativul obținut de un elev de clasa I sunt exemple de exprimări calitative ale unor caracteristici. În toate aceste situații vom putea utiliza aparatul statistic doar dacă vom face apel la operațiile de **cuantificare** şi **măsurare**.

Conform Dicționarului de Sociologie «Zamfir & Vlăsceanu (coord.), 1998, p.145», **cuantificarea** reprezintă *„operația teoretică de descriere cantitativă a fenomenelor şi proceselor sociale în vederea măsurării şi/sau evaluării acestora…*” În acelaşi sens, Mărginean (1982) face distincție între cuantificare, desfăşurată cu preponderență la nivel teoretico-metodologic şi **măsurare**, operație preponderent empirică, prin care se determină modalitatea de manifestare a fenomenului respectiv şi prin care se atribuie valori numerice unor caracteristici şi dimensiuni ale fenomenelor studiate.

Sintetizând o serie de considerații referitoare la cele două concepte, Luduşan şi Voiculescu (1997) consideră cunantificarea ca o operație complexă, ce implică trecerea de la conceptele abstracte la dimensiuni şi indicatori cantitativi, care, ulterior, prin acțiuni concrete să fie înregistrați şi, eventual, măsurați. **Cunatificarea**, susțin aceiaşi autori, este o operație prin care – pornindu-se de la analiza conceptelor ştiințifice, pe de o parte şi de la analiza naturii fenomenelor studiate, pe de altă parte – *„sunt dezvăluite şi definite componentele, dimensiunile şi expresiile cantitative ale domeniului cercetat, astfel încât să devină posibilă colectarea, înregistrarea şi exprimarea cantitativă a datelor şi folosirea aparatului statistico-matematic de analiză a acestora”* (p.22).

Mult mai contestat în ştiințele sociale, termenul de măsurare se referă la operația de atribuire de valori (sub formă de cifre sau simboluri) unităților statistice ale unei colectivități observate, pe baza unui set de reguli de atribuire a valorilor. Utilizarea acestor reguli este posibilă numai prin intermediul **instrumentelor de măsură**: termometru sau rigla, în cazul măsurării temperaturii sau lungimii; testul sau chestionarul, în cazul măsurării unor variabile psihologice sau sociologice. Odată

instrumentele construite, procesul de măsurare constă în citirea pe scalele acestor instrumente a unor valori reprezentând numărul de unități fundamentale de măsură. (Clocotici & Stan, 2001)

Scalele (nivelurile) de măsură nu sunt altceva decât regulile prin care sunt atribuite valori unităților statistice. „*Cunoaşterea proprietăților nivelurilor de măsură*, susține Mărginean (1982, p.70), *prezintă importanță deoarece s-a dovedit că o serie determinată de date permite, în mod legitim, să se adopte un anumit nivel de măsură sau tip de scală şi nu altul*."

Practica statistică, ținând cont de natura variabilelor şi, mai ales, de modul lor de exprimare (vezi cap. 2.2.), operează cu patru tipuri fundamentale de scale (niveluri de măsurare): scala nominală, scala, ordinală, scala de interval şi scala de raport. Fiecare dintre aceste scale se remarcă prin procedee specifice de exprimare numerică, ceea ce determină utilizarea anumitor operații de analiză şi prelucrare a datelor, foarte puține pentru nivelul nominal şi extrem de multe pentru cel de raport.

Încheiem prin a remarca unele proprietăți pe care trebuie să le îndeplinească o scală de măsură:

- să fie consistentă,
- să fie corectă,
- să fie exhaustivă şi
- să fie mutual exclusivă.

Scala are *consistență* internă dacă produce rezultate (aproape) identice, atunci când este folosită în mod repetat pentru acelaşi obiect sau fenomen; este *corectă* dacă produce informația pe care o aşteptăm de la ea; are proprietatea de a fi *exhaustivă* atunci când poate măsura toate entitățile cărora le este destinată; şi este *mutual exclusivă* atunci când, în urma măsurării, fiecare entitate primeşte o singură valoare (Clocotici & Stan, 2001).

2.4. SCALE DE MĂSURĂ

Scala nominală. Este cel mai simplu tip de scală şi presupune doar diferențierea calitativă a obiectelor şi fenomenelor măsurate. Aplicarea unei scale nominale la o colectivitate statistică înseamnă, în esență, o clasificarea a indivizilor după o caracteristică sau un atribut. Prin intermediul acestei scale se exprimă apartenența unităților statistice investigate la o categorie. Din aceste considerente, întâlnim acest tip de scală şi cu denumirile de scală calitativă, categorială sau de clasificare.

Condiția fundamentală ce se cere unei scale nominale este, de fapt, cerința elementară impusă oricărei clasificări: *dată fiind mulțimea claselor scalei şi mulțimea indivizilor, fiecare individ să se găsească în una şi numai una dintre clase* (Rotariu *et.al.*, 1999).

Un exemplu clasic de variabilă nominală utilizată în cercetările psiho-sociale este caracteristica „*gen*", ale cărei variante (categorii, atribute) sunt: *masculin* şi *feminin*. Chiar dacă, în activitatea concretă de înregistrare a datelor, celor două categorii le sunt atribuite codurile *1* şi *2* (la fel de bine putem codifica aceeaşi variabilă prin *m* şi *f*), aceste numere sunt doar nişte simboluri, între ele existând un

raport de echivalență şi nu unul de ordine. Nu putem afirma că *2* este „mai mult" decât *1*, ci doar că este diferit de acesta!

Alte scala nominale utilizate în psihologie şi sociologie sunt: - tipurilor temperamentale stabilite de Jung şi Eycenck: introvertit, extravertit, ambivert; - starea civilă: necăsătorit, căsătorit, văduv, …; opțiunea politică: partidul A, partidul B, …

Scala ordinală. Ca şi cea nominală, scala ordinală se foloseşte pentru exprimarea stărilor unor variabile calitative. În plus, acest tip de scală vine cu cerința ca *între categoriile (clasele) scalei să existe o relație de ordine*. Aceste scalele sunt cunoscute şi sub numele de scale de ordine, scale de rang sau scale ierarhice.

O scală ordinală permite *ordonarea* observațiilor, persoanelor, situațiilor de la mic la mare, de la simplu la complex etc., permițând astfel realizarea unor ierarhi (ranguri). În cazul scalelor ordinale putem stabili ierarhia celor „n" variante ale variabilei, însă nu putem preciza valoare diferenței dintre două variante.

Cel mai frecvent folosim acest tip de scală în studiul atitudinilor. Răspunsurile la o întrebare de genul *„Cât de mulțumit sunteți de relațiile din colectivul din care faceți parte*?" pot fi cuantificate printr-o scală ordinală, ale cărei clase sunt: *mulțumit, şi mulțumit şi nemulțumit, nemulțumit.*

Un alt exemplu de scală ordinală este ierarhia nevoilor umane în concepția psihologului american A. Maslow. Scala stabilită de el cuprinde următoarele categorii, ordonate de la simplu la complex: nevoi fiziologice; nevoi de securitate; nevoi sociale, de apartenență la grup; nevoia de stimă, de a fi apreciat de ceilalți; nevoia de autorealizare (Clocotici & Stan, 2001).

Clasele pot fi şi aici codificate prin cuvinte care să exprime semnificația lor sau prin simboluri. Dacă în cazul scalelor nominale simbolurile puteau fi atribuite oricum, de data aceasta ele trebuie să evidențieze ordinea claselor. Cel mai frecvent şi simplu mod de a evidenția ordinea este folosire numerelor naturale: 1, 2, 3 …. Atragem atenția că aceste simboluri numerice reprezintă *numere ordinale* şi nu cardinale, în consecință, operațiile aritmetice (adunarea, scădere, înmulțirea şi împărțirea) nu pot fi utilizate nici de această dată (Rotariu *et.al.*, 1999).

Scala de intervale. Împreună cu scalele de rapoarte, sunt utilizate pentru măsurarea variabilelor cantitative şi presupune atribuirea de valori numerice unităților colectivității. Din acest motiv ele se mai numesc *scări metrice* sau *numerice*.

Pe lângă cele două proprietăți impuse de nivelurile anterioare de măsurare, şi anume:

- *fiecare individ să se găsească în una şi numai una dintre clase,*
- *între categoriile (clasele) scalei să existe o relație de ordine,*

scalele metrice adaugă o a treia:

- *are sens luarea în considerare a distanțelor dintre categoriile scalei.*

Această proprietate face ca datele experimentale obținute pe o scală metrică să suporte aproape toate prelucrările statistice posibile.

Caracteristic pentru scala de interval este faptul că utilizează o valoare 0 convențională. Astfel, măsurarea cu acest tip de scală este independentă de originea aleasă şi de unitatea de măsură folosită, putându-se trece de la un sistem de măsurare la altul.

Exemplul clasic îl reprezintă măsurarea temperaturii în sistemul Celsius şi în sistemul Fahreinheit. Trecând de la un sistem de măsurare la altul, deci schimbând zeroul convenţional şi valorile temperaturii, raportul dintre două modificări de temperatură rămâne acelaşi (Jaba & Grama, 2004). Un alt exemplu de astfel de scală îl reprezintă scalele pentru măsurarea inteligenţei.

Referindu-se la proprietăţile scalelor de interval, M. Popa (2004) atrage atenţia asupra faptului că valorile obţinute prin măsurări de acest tip nu ne permit evaluări de genul: „O temperatură de 10 grade este de două ori mai mare decât una de 5 grade" sau, „O persoană care a obţinut un scor de 30 de puncte este de două ori mai inteligentă decât una care a obţinut 15 puncte". Aceasta, deoarece nici temperaturile măsurate pe scala Celsius şi nici inteligenţa nu au o valoare 0 absolută (dacă acceptăm că nici un om viu nu are inteligenţă nulă).

De asemenea, trebuie remarcat faptul că cele mai multe dintre variabilele psihologice sunt expresia unor evaluări subiective, aspect ce face greu de demonstrat egalitatea intervalelor dintre două valori consecutive. Uneori, chiar şi în cazul unor măsurători extrem de exacte este dificil de asumat acest lucru. De exemplu, dacă măsurăm „iubirea" la un eşantion de cupluri care se plimbă, prin durata „ţinerii de mână", nu putem fi siguri că diferenţa de „iubire" dintre cei care se ţin de mână 10 minute şi cei care se ţin de mână 20 de minute este aceeaşi ca în cazul diferenţei dintre 20 şi 30 de minute. Cu toate acestea, multe dintre măsurătorile studiilor psihologice sunt asimilate scalei de tip interval. (Popa, 2004)

Scala de rapoarte sau scala de proporţii (sau scala de interval cu origine raţională). Face parte din categoria scalelor metrice, fiind folosită tot pentru exprimarea variabilele cantitative.

Această scală de măsură posedă ca note distinctive existenţa unei origini naturale (a unui 0 absolut; altfel spus, nu există nici o valoare mai mică decât valoarea 0) şi precizarea clară a semnificaţiei unităţii de măsură, ceea ce face posibilă compararea raporturilor dintre gradaţiile scalei.

Scala de rapoarte se foloseşte pentru măsurarea valorilor unor variabile precum venitul, înălţimea, timpul de reacţie ş.a.

După uni autori (Kinnear şi Gray, 2000, *cf*. Sava, 2004a) şi după cum reiese şi din utilizarea programului SPSS, în care există doar trei niveluri de măsurare, tendinţa actuală este de a renunţa la diferenţierea între ultimele două tipuri de scale. Aceasta pentru că majoritatea procedurilor statistice utilizate în cazul scalelor de intervale sunt valabile şi pentru scalele de rapoarte. Termenul generic sub care se reunesc cele două tipuri de scale este cel de scală numerică sau metrică.

2.5. DEFINIREA VARIABILELOR STATISTICE CU AJUTORUL SPSS

Pentru **crearea unei baze de date** se începe prin definirea variabilelor. După apariţia ferestrei de întâmpinare din editorul de date SPSS se deschide fereastra de gestionare a variabilelor unde, pentru fiecare variabilă, sunt specificate următoarele caracteristici:

- Name – numele variabilei (*de exemplu*: sex).
- Type – tipul variabilei, poate fi numeric, dată calendaristică, string ş.a. (*în exemplul nostru*: numeric).
- Width – numărul de caractere al variabilei (*ex.*: 1).
- Decimals – pentru variabilele numerice trebuie specificat numărul de caractere după virgulă al variabilei (*ex.*: 0).
- Label – comentariu (eticheta) ce însoţeşte variabila (*ex.*: sexul subiectului).
- Values – valorile pe care le poate lua variabila şi comentariile/etichetele ataşate acestora (*ex.*: 1 = „masculin”; 2 = „feminin”).
- Missing – specificarea cazurilor omise (*ex.*: None).
- Columns – numărul de spații alocat în editorul de date acestei variabile (ex.: 8).
- Align – alinierea acestei variabile în editorul de date, poate fi aliniere la stânga, la dreapta sau centrat (*ex.*: Center).
- Measure – Nivelul de măsurare al variabilei (tipul scalei), poate fi numeric (scale), ordinal şi nominal (*ex.*: Nominal).

3.

ORDONAREA, GRUPAREA ŞI PREZENTAREA DATELOR STATISTICE

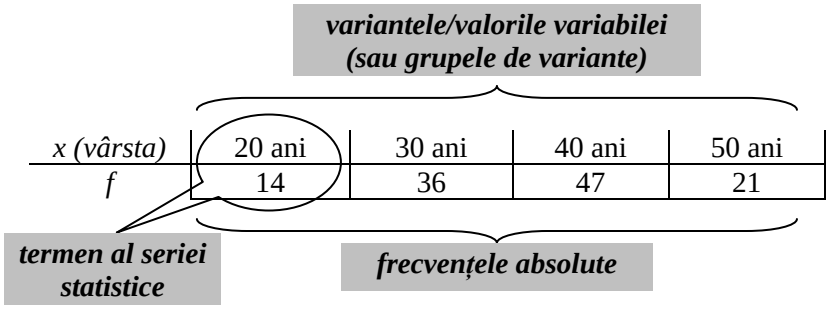
3.1. Serii (distribuții) statistice
3.2. Gruparea (sistematizarea) datelor
3.3. Prezentarea datelor sub formă de tabele
3.4. Reprezentarea grafică a datelor statistice
3.5. Utilizarea SPSS pentru ordonarea şi gruparea datelor statistice
3.6. Utilizarea SPSS pentru prezentarea datelor statistice sub formă de tabele
3.7. Utilizarea SPSS pentru reprezentarea grafică a datelor statistice

3.1. SERII (DISTRIBUȚII) STATISTICE

În cazul unui număr foarte mare de date este imposibilă (şi inutilă) analiza fiecărei valori în parte. În această situație, înaintea prelucrării şi analizei datelor se procedează la ordonarea, gruparea şi organizarea lor. Rezultatul ordonării şi grupării datelor statistice îl constituie ***seriile (distribuțiile) statistice de frecvențe***.

Acestea sunt formate din două şiruri paralele de date din care unul reprezintă variantele/valorile variabilei (sau grupele de variante) iar celălalt numărul de unități statistice corespunzătoare fiecărei valori sau variante (frecvențele absolute sau relative). Fiecare frecvență asociată valorii/variantei respective a caracteristicii studiate reprezintă un termen al seriei statistice.

Exemplu:



În funcție de modul de prezentare al variantelor, seriile statistice, se împart în:

serii simple – obținute prin simpla înşiruire a valorilor individuale. Acestea sunt ulterior supuse operațiilor de ordonare şi grupare (dacă numărul lor este suficient de mare), obținându-se astfel unul din următoarele două tipuri de serii.

serii de (pe) variante – când fiecărei variante îi revine un anumit număr de unități.
serii de (pe) intervale – când fiecărui interval, mărginit de o limită inferioară şi de una superioară, îi revine un anumit număr de unități.

Ultimele două tipuri se mai numesc şi ***serii (repartiții) de frecvențe*** şi formează ceea ce numim o ***DISTRIBUȚIE STATISTICĂ***.

În funcție de natura şi modul de manifestare ale variabilei studiate distingem două tipuri principale de serii statistice: serii statistice *cantitative* sau *calitative*. La acestea putem adăuga alte două tipuri de distribuții statistice, la care criteriul după care se face diferențierea este spațiul sau timpul: serii statistice *spațiale* şi *cronologice*.

Aceste criterii nu numai că realizează o clasificare a seriilor statistice dar, vom vedea în capitolele următoare, determină limitele şi specificul prelucrărilor statistice complexe. Atunci când variabilele sunt cantitative vom vorbi despre ***tehnici statistice parametrice***; în celălalt caz, al caracteristicilor calitative, prelucrările ce le vom efectua vor fi de tip ***non-parametric***.

În concluzie, seria statistică de frecvențe este rezultatul operațiilor de ordonare şi grupare. Prezentarea seriilor statistice se face sub forma înşiruirii, pe orizontală sau pe verticală, a unor perechi de numere sau expresii, în care primul element reprezintă caracteristica (ce poate fi cantitativă sau calitativă, spațială sau cronologică), iar al doilea frecvența, întotdeauna numerică, a variantelor sau grupelor de variante ce delimitează caracteristica respectivă. În rapoartele de cercetare aceste distribuții statistice, unele reflectând mai multe caracteristici concomitent, sunt ilustrate cu ajutorul tabelelor şi al graficelor.

Reamintim următoarele notații cu care operăm în prezentarea şi prelucrarea distribuțiilor statistice:

- variantele sau grupele (clasele) de variante, x_i: $x_1, x_2, \dots x_k, \dots$
- frecvența variantei x_i (numărul de apariții), f_i: $f_1, f_2, \dots f_k, \dots$
- numărul total de variante (total frecvențe) n: $n = \Sigma f_i \quad i = 1, 2, \dots k, \dots$

În cazul seriilor statistice de intervale se presupune că toate valorile din interiorul fiecărei grupe (clase) se concentrează în *valoarea centrală a clasei*, notată tot cu x_i. Această valoare va înlocui în seria statistică intervalul respectiv şi se calculează ca medie aritmetică a valorilor extreme ale intervalului:

$$x_i = \frac{x_{max} + x_{min}}{2} \qquad (3.1)$$

Menționăm faptul că o distribuție statistică poate reda pe lângă frecvențele absolute (f sau f_a) şi pe cele relative (f_r). Acestea sunt absolut necesare când se doreşte compararea unor eşantioane cu numărul total de variante (n) diferit (*de exemplu*: în cazul a două clase cu număr total de elevi diferit). Mai mult, atunci când prelucrările statistici ulterioare o impun, putem determina şi alte frecvențe:

- frecvența (absolută sau relativă) cumulată crescător, dată de suma frecvențelor valorilor care apar până la valoarea x_i respectivă, inclusiv;
- frecvența (absolută sau relativă) cumulată descrescător, dată de suma frecvențelor valorilor care apar de la valoarea x_i respectivă, inclusiv.

3.2. GRUPAREA (SISTEMATIZAREA) DATELOR

Gruparea statistică reprezintă o operație de sistematizare a populației pe părți statistic omogene în funcție de variația[1] unei variabile (sau a mai multora).

Importanța acestei operații inițiale derivă din erorile ce pot fi induse fie în cazul stabilirii unui număr foarte mare de grupe (clase) – situație în care se ajunge la „fărâmițarea" colectivității –, fie în situația alegerii unui număr prea mic de grupe, cu intervale foarte mari în cadrul lor – situație în care nu vom surprinde tipurile calitative existente.

În cazul variabilelor numerice (cantitative) putem realiza

1) <u>grupări pe variante</u> – utilizate în cazul variabilelor de tip discret, când ele pot lua doar valori întregi (*exemple*: numărul membrilor unei familii, notele şcolare).
2) <u>grupări pe intervale</u> – utilizate în cazul variabilelor de tip continuu, când ele pot lua orice valoare într-un interval finit sau infinit (*exemple*: timpul de reacție, mediile şcolare anuale, înălțimea).

Menționăm faptul că şi variabilele de tip discret pot fi supuse grupărilor pe intervale (*exemplu*: note între 2 şi 4; 5–7; 8–10 etc.). În ambele situații *mărimea intervalului* (K) se obține cu ajutorul formulei lui H.A. Sturges:

$$K = \frac{x_{max} - x_{min}}{1 + 3{,}322 \bullet \lg n} \qquad (3.2)$$

unde, n reprezintă numărul total de variante.

În situația în care numărul de grupe este ales de cercetător (bazându-se pe experiență şi intuiție), mărimea intervalului (K) rezultă astfel:

- în cazul variabilelor de tip continuu, prin raportarea *amplitudinii variației* ($A = x_{max} - x_{min}$) la numărul de grupe:

$$K = \frac{x_{max} - x_{min}}{\text{nr. grupelor}} \qquad (3.3)$$

- în cazul variabilelor de tip discret, prin raportarea *numărului valorilor diferite ale variabilei* ($N_x = x_{max} - x_{min} + 1 = A + 1$) la numărul de grupe:

$$K = \frac{x_{max} - x_{min} + 1}{\text{nr. grupelor}} \qquad (3.4)$$

[1] Variația reprezintă proprietatea unei variabile de a înregistra mai multe valori (în cazul variabilelor cantitative) sau mai multe forme de manifestare (în cazul variabilelor calitative) (Blezu, 2002).

O atenție deosebită trebuie acordată precizării limitelor sau capetelor intervalelor. În cazul caracteristicilor discrete limitele intervalelor ies foarte bine în evidență, ele fiind diferite (*exemplu*: intervalele 2–4; 5–7; 8–10).

Mai delicat este cazul caracteristicilor continui, când trebuie precizat care dintre intervale include limita sau, altfel spus, care capăt al intervalului este deschis/închis (*exemplu*: intervalele (2–4]; (4–6]; (6–8] etc. sunt deschise în partea stângă). Pentru evitarea confuziilor se procedează din start la departajarea limitelor, astfel: 2,01–4; 4,01–6; 6,01–8 etc.

3.3. PREZENTAREA DATELOR SUB FORMĂ DE TABELE

Prezentarea datelor sub forma unui tabel statistic permite atât o bună vizualizare cât şi, mai ales, efectuarea diverselor calcule în procesul de prelucrare a datelor.

În elaborarea unui tabel pot fi identificate următoarele elemente şi reguli principale (Novak, 1995):

- titlul tabelului - care trebuie să fie clar, scurt şi să definească exact fenomenul pe care îl reprezintă şi, după caz, perioada la care se referă;
- macheta tabelului - formată din liniile orizontale (rânduri) şi liniile verticale (coloane) din întretăierea cărora apar rubricile (celulele, căsuțele) care conțin datele numerice şi/sau denumirile textuale;
- subiectul tabelului - înscris de obicei la capătul rândurilor, este constituit din unitățile populației statistice (ex.: grupe de note, grupe de puncte etc);
- predicatul tabelului - înscris de obicei la capătul coloanelor, cuprinde ansamblul indicatorilor care se înregistrează la nivelul unităților populației statistice;
- indicarea obligatorie a sursei de date, atunci când este cazul (de obicei sub tabel);
- se recomandă indicarea unităților de măsură în care se exprimă datele (de obicei, între titlul şi macheta tabelului);
- se recomandă numerotarea tabelelor - pentru identificarea mai uşoară a acestora în textul de analiză.

În funcție de scopul întocmirii, de conținutul lor şi de numărul caracteristicilor studiate tabelele pot fi de mai multe tipuri. Astfel:

a) Tabele ale unor serii statistice

Pot fi întocmite atât pentru seriile de variante cât şi pentru cele de intervale. Diferența este dată de rândurile tabelului care vor constitui variantele seriei, în primul caz, sau clasele de variante (eventual valorile centrale), în cel de-al doilea caz. În ambele situații pe coloane vor fi trecute frecvențele, absolute sau relative, cumulate sau descrescătoare. (*Exemplu*: a se vedea tabelul 3.3)

b) Tabele centralizatoare

Sunt utilizate în toate situațiile în care un număr mare de date trebuie stocate şi conservate în vederea prelucrării lor ulterioare. În lucrările ştiințifice aceste tabele sunt, de obicei, prezentate sub formă de anexe, şi conțin pe coloane totalitatea variabilelor studiate, iar pe rânduri, totalitatea unităților statistice (colectivitatea statistică) investigate.

c) Tabele comparative

Cuprind fie datele obținute pe eşantioane diferite pentru aceeaşi caracteristică, fie datele aceluiaşi eşantion pentru caracteristici diferite.

d) Tabele cu dublă sau triplă intrare

În acest caz, şi coloanele şi rândurile exprimă variațiile uneia sau a două caracteristicii (variabile). Fiecare celulă exprimă numărul de unități statistice caracterizate prin variantele corespunzătoare tuturor caracteristicilor de pe orizontală şi verticală.

3.4. REPREZENTAREA GRAFICĂ A DATELOR STATISTICE

Cu ajutorul reprezentărilor grafice sunt vizualizate informațiile statistice, facilitându-se perceperea pe ansamblu a datelor, sesizarea unor aspecte privind variația valorilor observate, repartiția lor, legăturile existente între ele ş.a.

Graficul trebuie să cuprindă:

- titlul - care poate fi plasat fie sub, fie deasupra graficului şi trebuie să precizeze limpede fenomenul pe care îl reprezintă;
- legenda – utilizată pentru specificarea anumitor simboluri sau convenții utilizate;
- sistemul axelor rectangulare (dacă este cazul) - în care linia orizontală (abscisă) cuprinde valorile variabile x, iar cea verticală (ordonată) cuprinzând frecvențele f;
- se recomandă numerotarea graficelor - pentru identificarea mai uşoară a acestora.

Graficele cel mai des utilizate sunt ***graficele de tip bară***, ***histogramele, poligoanele de frecvențe***, şi ***curbele de distribuție***, pe abscisă notându-se intervalele de variație (sau variantele), iar pe ordonată frecvențele corespunzătoare acestor intervale (sau variante). Aceste reprezentări grafice se obțin prin unirea intersecțiilor perpendicularelor ridicate din punctele perechi de pe cele două axe. În cazul seriilor de intervale perpendiculara pentru desemnarea valorii frecvenței se ridică din mijlocul intervalului, respectiv din punctul corespunzător valorii centrale a clasei.

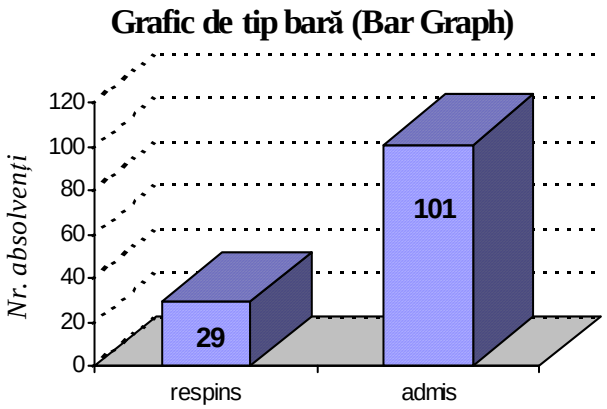
Graficele de tip bară[2] le folosim când dorim să reprezentăm fie variabile cantitative discrete, fie variabile categoriale (măsurate prin scale nominale sau ordinale). Caracteristic acestui tip de grafic este faptul că barele verticale sunt delimitate de un spațiu, iar ordinea barelor poate fi schimbată.

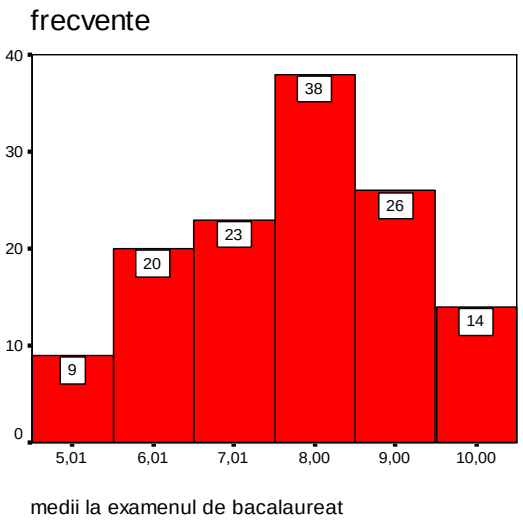
Histogramele şi poligoanele de frecvențe sunt reprezentările grafice utilizabile în cazul seriilor statistice cantitative, însă numai atunci când variabilele sunt continue. De exemplu, situația absolvenților de liceu după examenul de admitere la facultate (exprimată prin două variante: „admis”, „respins”) va fi reprezentată printr-un grafic de tip bară (deoarece avem de-a face cu o variabilă calitativă, măsurată printr-o scală

[2] În engleză: *bar graph*.

nominală), iar mediile la bacalaureat ale aceloraşi absolvenţi printr-o histogramă sau printr-un poligon de frecvenţe (deoarece avem o variabilă cantitativă continuă sau, altfel spus, o variabilă măsurată printr-o scală numerică).

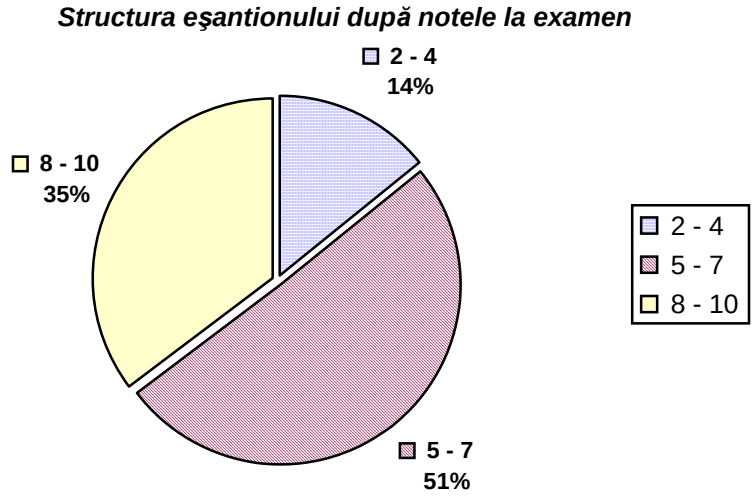
Pentru a evidenţia şi/sau compara structurile se utilizează ***diagramele de structură***, construite cu ajutorul suprafeţelor (cercuri, pătrate, dreptunghiuri), ***diagramele de comparaţie*** şi ***reprezentările prin figuri simbolice*** ş.a.. În multe cazuri, sunt studiate mai multe caracteristicii folosindu-se reprezentări grafice complexe precum: ***piramide ale vârstelor***, ***grafice comparative***, ***grafice combinate***.





În ce priveşte diagramele sub forma figurilor geometrice (cerc, pătrat, dreptunghi) utilizate atât pentru prezentarea structurilor cât şi/sau pentru compararea în timp a evoluției fenomenelor se procedează astfel (Novak, 1995):

- se construiesc cele două figuri în aşa fel, încât raportul dintre raze (sau laturi) să fie proporțional cu nivelurile fenomenului studiat în cele două perioade diferite de timp (în două localități etc.);
- în cadrul fiecărei figuri geometrice se reprezintă structura corespunzătoare anului (spațiului geografic) respectiv.



3.5. UTILIZAREA SPSS PENTRU ORDONAREA ŞI GRUPAREA DATELOR STATISTICE

ORDONAREA DATELOR STATISTICE CU AJUTORUL SPSS

Se parcurge, în bara de meniuri, traseul:

„Data" – „Sort cases..."

Va fi afişată fereastră de dialog din figura 3.1.

După ce selectăm variabila după care dorim să facem ordonarea (prin trecere ei din stânga în fereastra intitulată „Sort by:") ne mai rămâne să alegem sensul ordonării: crescător/ascendent sau descrescător/descendent. Se poate realiza sortarea datelor după mai multe variabile; în acest caz, se va ține cont de ordinea variabilelor în fereastra „Sort by:".

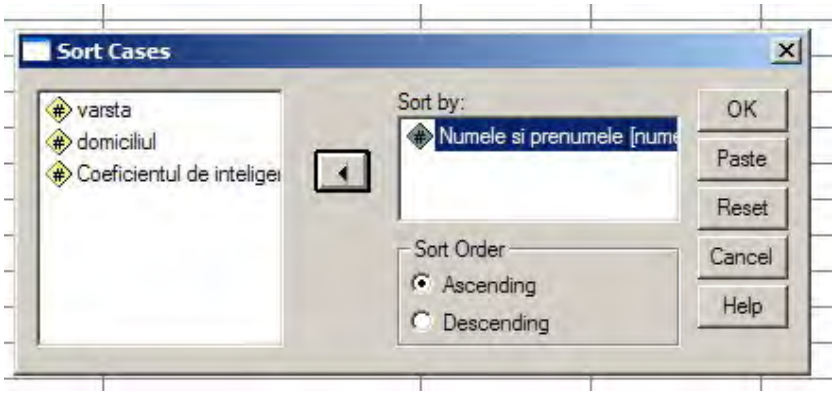


Figura 3.1. *Fereastră de dialog pentru sortarea (ordonarea) datelor*

3.6. UTILIZAREA SPSS PENTRU PREZENTAREA DATELOR STATISTICE SUB FORMĂ DE TABELE

Pentru calcularea frecvențelor absolute şi/sau relative ale unei serii statistice simple sau de variante, precum şi pentru redarea sub formă tabelară a distribuției de frecvențe, se parcurge, în bara de meniuri, traseul:

„Analyze" – „Descriptive Statistics" – „Frequencies…"

Vom fi întâmpinați de fereastra următoare, în care, în partea stângă sunt afişate toate variabilele din baza de date (în ordine alfabetică sau în ordinea definirii lor).

Figura 3.4. *Fereastra de întâmpinare (de dialog) pentru calculul frecvențelor*

3.7. UTILIZAREA SPSS PENTRU REPREZENTAREA GRAFICĂ A DATELOR STATISTICE

Pentru a obține o reprezentare grafică aferentă seriei statistice respective, revenim la fereastra de întâmpinare pentru calculul frecvențelor (figura 3.4) şi apăsăm butonul „*Charts…*”.

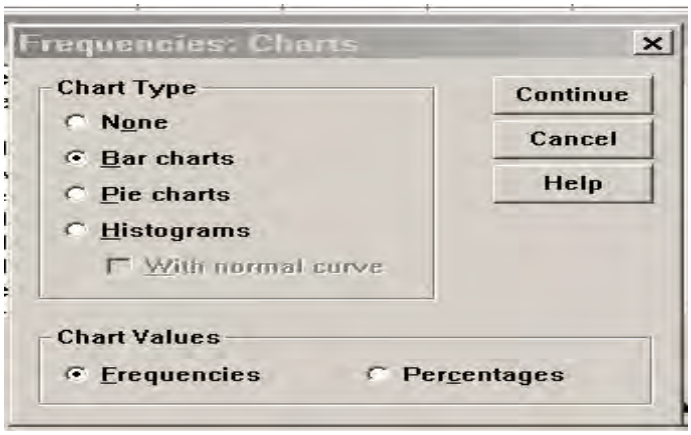


Figura 3.7. *Fereastră de opțiuni pentru reprezentarea grafică a datelor statistice*

Va apărea o nouă fereastră în care, înainte de a apăsa butonul „*Continue”*, vom opta pentru una din următoarele situații („*Chart Type”*):

- „*None”*, când nu se doreşte reprezentarea grafică a variabilei;
- „*Bar charts”*, reprezentare (printr-un „grafic de tip bară”) folosită pentru serii statistice calitative sau pentru seriile de frecvențe (de variante sau de intervale) în care variabila este discontinuă; aici putem opta pentru afişarea valorilor pe grafic („*Chart Value”*) sub forma frecvențelor absolute („*Frequencies*”) sau a celor relative („*Percentages*”);
- „*Pie charts”*, reprezentare grafică sub forma diagramei de structură prin arce de cerc folosită pentru serii statistice calitative sau pentru seriile de frecvențe (de variante sau de intervale) cu număr redus de variante; avem posibilitatea aceloraşi opțiuni de mai sus;
- „*Histograms”*, reprezentare grafică sub formă de histogramă folosită pentru serii statistice cantitative simple sau de variante, în care variabila este de tip continuu; aici se poate opta pentru trasarea curbei distribuției normale prin activarea căsuței „*With normal curve*”.

4.

INDICATORI AI TENDINȚEI CENTRALE

4.1. Mediile
4.2. Quantilele: mediana, quartilele, decilele şi centilele
4.3. Modul
4.4. Relația dintre indicatorii tendinței centrale
4.5. Reprezentări de tip *Boxplots*
4.6. Utilizarea SPSS pentru calcularea şi reprezentarea indicatorilor de poziție

În cele mai multe investigații psihosociale sau educaționale prezentarea rezultatelor sub formă tabelară sau prin reprezentări grafice nu este suficientă. Prin intermediul unor indicatori statistici putem realiza o prelucrare mult mai riguroasă a datelor, putem cunoaşte mult mai temeinic fenomenele studiate.

Termenul de „**indicator**" se referă la acele „*valori ataşate variabilelor statistice cantitative prin intermediul cărora se încearcă exprimarea, de o manieră sintetică a informației conținută în distribuția de frecvențe respectivă*" (Rotariu *et. al.*, 1999, p. 42).

În funcție de natura informației oferită de indicatorii statisticii, aceştia se clasifică în trei mari categorii:

- indicatori ai tendinței centrale (de poziție sau de nivel),
- indicatori ai variației (de dispersie sau de împrăştiere),
- indicatori ai formei distribuției.

Pentru a determina modul în care datele statistice tind să graviteze în jurul unor valori centrale se folosesc ***indicatorii tendințelor centrale***. Dintre aceştia vom prezenta: media, quantilele (mediana, quartilele, decilele şi centilele) şi modul.

4.1. MEDIILE

Mărimile medii exprimă ceea ce este comun şi general în forma de manifestare a fenomenelor studiate.

Pentru a ne fi de folos, însă, calculul mărimilor medii trebuie să îndeplinească anumite condiții:

- să se bazeze pe un număr suficient de mare de cazuri individuale;
- valorile individuale ale caracteristicii să nu difere prea mult de la o unitate statistică la alta, adică să avem o colectivitate omogenă;

- mărimea medie aleasă pentru calcul să corespundă cel mai bine formei de variație a caracteristicii studiate şi să valorifice cel mai bine materialul cifric de care dispunem (Novak, 1995).

MEDIA ARITMETICĂ

Media aritmetică **(m,** $\bar{x}$ sau **μ[1])**, reprezintă, în cazul datelor negrupate (serii simple), raportul dintre suma valorilor variabilei respective şi numărul lor.

$$m = \frac{\Sigma x_i}{n} \qquad (4.1)$$

Dacă datele sunt grupate (distribuții de frecvențe), media - numită uneori medie aritmetică ponderată[2] - va fi:

$$m = \frac{\Sigma x_i \cdot f_i}{\Sigma f_i} \qquad (4.2)$$

În cazul grupării valorilor pe intervale, în formula de mai sus x_i reprezintă valoarea centrală a intervalului.

Proprietățile mediei aritmetice:

- dacă la toate valorile seriei statistice se adaugă (scade) o constantă *c*, atunci media se măreşte (scade) cu acea valoare: dacă $y_i = x_i + c$, atunci $m_x = m_y + c$
- dacă toate valorile seriei statistice se înmulțesc (divid) cu o constantă *c*, atunci şi media se va multiplica (divide) cu aceeaşi valoare *c*: dacă $y_i = c \cdot x_i$, atunci $m_y = c \cdot m_x$
- suma abaterilor valorilor de la medie este întotdeauna nulă: $\sum x_i - m = 0$
- suma pătratelor abaterilor de la medie va fi întotdeauna mai mică decât suma pătratelor abaterilor de la oricare alt punct al distribuției.

4.2. QUANTILE[3]

O altă categorie de indicatori ai tendințelor centrale o reprezintă ***quantilele***. Acestea sunt indicatori de poziție şi au rolul de a împărții seria de date într-un anumit număr de părți. Dintre quantilele cele mai des calculate amintim:

---

[1] m şi $\bar{x}$ (x barat) se folosesc atunci când ne referim la media unui eşantion (situația cea mai frecventă), iar μ (miu) atunci când calculăm media întregii populații de referință.

[2] Pentru a înțelege corect sensul termenului de *medie ponderată* recomandăm următoarea referință bibliografică: Rotariu *et. al.*, 1999, pp. 43-44.

[3] În limba engleză, se numesc *percentiles*.

***Mediana* (M sau M_e),** este valoarea care împarte seria ordonată de date în două părți egale. Jumătate din valori (50%) se găsesc în partea stângă a medianei iar cealaltă jumătate în partea dreaptă.

Pentru calculul medianei este absolut necesară ordonarea seriei statistice, fie crescător, fie descrescător (aspect fără importanță în cazul calculului valorilor medii!).

Pentru a afla al câtelea element al unei serii cu <u>număr impar</u> de termeni este mediana se calculează cota medianei după formula;

$$\text{Cota } M = (n+1)/2 \qquad (4.7)$$

De exemplu, presupunând că notele, ordonate crescător, obținute de un lot de nouă subiecți sunt:

4 5 6 7 **<u>7</u>** 8 8 8 9

cota medianei va fi (9+1)/2 = 5, astfel încât mediana va corespunde celui de-al cincilea termen din serie, adică 7. Se observă că şi în stânga şi în dreapta acestei valori se află un număr egal de termeni.

Pentru seriile formate dintr-un <u>număr par</u> de valori formula (4.7) rămâne valabilă, numai că rezultatul nu va mai fi întotdeauna un număr întreg. Vom vorbi de doi termeni centrali, poziția medianei fiind între termenul n/2 şi (n/2)+1. În acest caz, mediana se calculează făcând media celor două valori, putând să coincidă (dacă valorile corespunzătoare termenilor n/2 şi (n/2)+1 sunt egale), sau nu (în caz contrar), cu una din valorile seriei.

Dacă în exemplu anterior mai apare un subiect cu nota 9 vom avea o serie cu zece termeni:

4 5 6 7 **<u>7</u>** **<u>8</u>** 8 8 9 9

mediana va fi dată de media valorilor corespunzătoare termenilor cinci şi şase, adică 7,5.

Lucrurile devin mult mai complicate dacă ne referim la distribuții de frecvențe[4].

***Quartilele* (Q)** reprezintă alte tipuri de quantile, ele împărțind seria de date în patru părți egale, astfel:

- quartila 1 (Q_1) împarte valorile în 25% (un sfert) şi, respectiv, 75% (trei sferturi);
- quartila 2 ($Q_2 = M$) împarte seria de date în două jumătăți egale, ea fiind, de fapt, mediana;
- quartila 3 (Q_3) împarte seria ordonată în 75% şi, respectiv, 25%.

[4] Pentru unii indicatori ai tendinței centrale formulele de calcul sunt mai complexe atunci când datele sunt grupate. Tratatele de statistică aplicată prezintă în amănunt toate aceste formule.

Analog, se definesc şi celelalte quantile: ***decilele*** (împart o serie ordonată în zece părți egale) şi ***centilele*** (împart o serie ordonată într-o sută de părți egale).

4.3. MODUL (VALOAREA MODALĂ)

***Modul* «sau *valoarea modală*» (M_o),** reprezintă valoarea caracteristicii care prezintă frecvența cea mai mare, care apare de cele mai multe ori în seria de date.

De exemplu, în cazul unei serii simple de date de forma:

4 5 5 6 7 7 **8 8 8** 9

modul va fi 8, această valoare apărând de cele mai multe ori în cadrul seriei.

Pentru o serie de variante, modul este egal cu varianta care are cea mai mare frecvență, iar pentru o serie de intervale, fie se calculează media intervalului cu cea mai mare frecvență, fie rămânem doar la noțiunea de ***interval modal***.

De cele mai multe ori seriile statistice au un singur mod, situație în care spunem că avem o distribuție *unimodală*. Dacă întâlnim două sau mai multe valori modale vom avea distribuții *bi-* sau *multimodale* (vezi capitolul 6.3.).

4.4. RELAȚIA DINTRE MEDIE, MEDIANĂ ŞI MODUL

În funcție de aspectul (grafic) al unei serii statistice cele trei valori medii pot să coincidă, sau nu. În prima situație vom vorbi de o distribuție normală (gaussiană) sau vom afirma că populația din eşantionul studiat este distribuită „normal", este omogenă în raport cu variabilă respectivă (vezi capitolul 6.3.).

În celălalt caz, nu toți cei trei indicatori sunt reprezentativi; va trebui să ținem seama de modul de exprimare al variabilei, motiv pentru care se impun următoarele precizări:

- media este recomandată în cazul variabilelor numerice care îndeplinesc condițiile parametrice (distribuție normală, omogenitate ş.a.);
- mediana se recomandă pentru cazurile în care nu sunt îndeplinite condițiile parametrice (distribuții asimetrice, eterogenitate crescută etc) şi în cazul variabilelor de tip ordinal
- modul este utilizat mai rar pentru date numerice, fiind însă foarte util în cazul variabilelor de tip categorial (date calitative, nominale), deoarece nu putem calcula ceilalți parametrii centrali (Sava, 2004b).

Între aceste trei caracteristici medii de bază există o relație aproximativă, stabilită de G.U. Yule şi M.G. Kendall, valabilă pentru distribuții moderat asimetrice:

$$M_o = M_e - 3(m - M_e) \qquad (4.8)$$

4.5. REPREZENTĂRI TIP *BOXPLOT*

O modalitate specifică de a reprezenta tendința cazurilor unei serii statistice de a se grupa în jurul unor valori centrale o reprezintă diagramele de tip ***Boxplot***. Acestea marchează printr-un dreptunghi (o cutie) cele trei quartile – Q1, Q2, și Q3 – ale oricărei serii statistice şi prin două linii distincte cea mai mică, respectiv cea mai mare valoare a seriei. Din acest motiv, despre această reprezentare se mai spune că reprezintă o ***rezumare prin cinci valori***.

Între cele două quartile Q1 și Q3 (în interiorul dreptunghiului) se regăsesc 50% din cazuri. Mai mult, sunt reprezentate, atunci când este cazul, **valorile extreme**[5] (mai mici/mari de 1.5, respectiv 3 lungimi de cutie[6] – simbolizate prin cerc, respectiv asterisc).

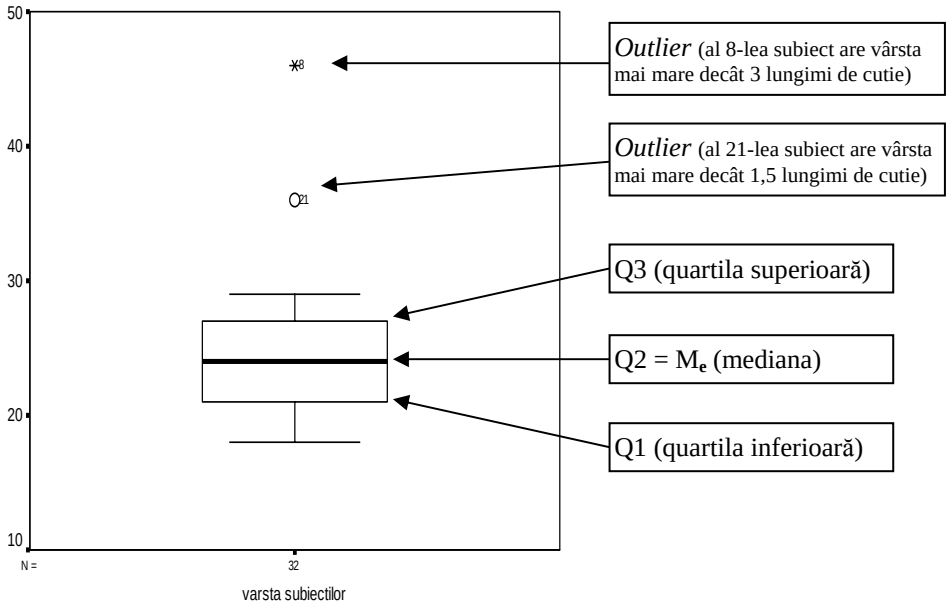


Figura 4.1. *Reprezentare grafică de tip* Boxplot *a variabilei „Vârsta subiecților"*

[5] În engleză, *outliers*.

[6] Lungimea (înălțimea) cutiei reprezintă abaterea interquartilă: $I = Q_3 - Q_1$ - vezi cap. 5.1.

4.6. UTILIZAREA SPSS PENTRU CALCULAREA ŞI REPREZENTAREA GRAFICĂ A INDICATORILOR DE POZIȚIE

Cu ajutorul programului SPSS valorile tendinței centrale se obțin cu mare ușurință, existând mai multe posibilități.

Una dintre posibilități este amintită în capitolul anterior, presupunând traseul:

„Analyze" – „Descriptive Statistics" – „Frequencies…"

După ce, în fereastra de dialog pentru calculul frecvențelor (vezi figura 3.4.), selectăm variabila sau variabilele dorite, apăsăm butonul „*Statistics…*" și vom pătrunde într-o nouă fereastră de opțiuni (figura 4.2).

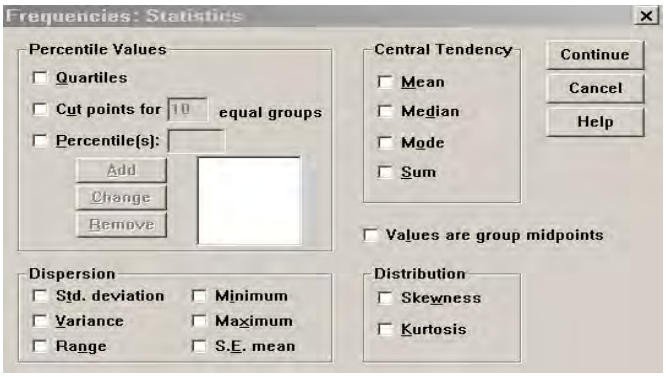


Figura 4.2. *Fereastră de opțiuni pentru calculul unor indicatori statistici*

La rubrica „*Percentile Values*" putem opta pentru calculul quartilelor sau a oricăror altor quantile (*Percentiles*) care să împartă seria în intervale egale (*equal groups*), sau inegale.

La rubrica „*Central Tendency*" se optează pentru calcularea mediei aritmetice (*Mean*), medianei (*Median*), Modului (*Mode*) sau sumei valorilor (*Sum*).

5.

INDICATORI AI VARIAȚIEI ŞI INDICATORI AI FORMEI

5.1. Indicatori simpli (elementari) ai variației
5.2. Indicatori sintetici ai variației
5.3. Indicatori ai formei distribuției
5.4. Utilizarea SPSS pentru calcularea indicatorilor variației şi ai formei

Utilizarea mediei pentru caracterizarea a ceea ce este comun şi tipic în colectivitățile statistice trebuie să fie însoțită de verificarea reprezentativității acesteia pentru întreaga serie de valori individuale. Vom analiza cu ajutorul unei alte categorii de indicatori, numiți **indicatori ai variației** (**de dispersie** sau **de împrăştiere**), măsura în care valorile individuale variază în jurul mediei sau, altfel spus, gradul de împrăştiere (de dispersie) a indivizilor în cadrul seriei de valori pe care aceştia le iau. Putem avea serii statistice cu aceeaşi medie, însă cu o distribuție a valorilor diferită, adică eşantioane diferite din punct de vedere al variabilității şi omogenității (vezi figura 5.1.).

La rândul lor, indicatorii variației se împart în indicatori simpli şi indicatori sintetici.

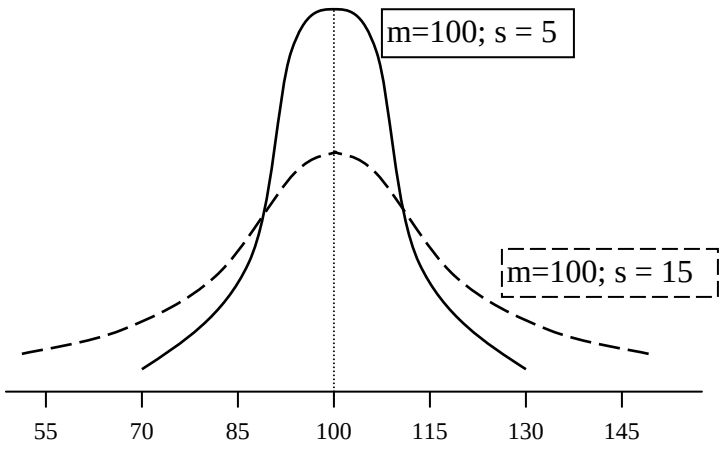


Figura 5.1. *Distribuții statistice cu aceleaşi valori centrale, dar cu grade diferite de variabilitate*

5.1. INDICATORI SIMPLI (ELEMENTARI) AI VARIAȚIEI

Se obțin prin compararea a doi termeni din serie sau prin compararea oricărui termen al seriei cu o valoare fixă din cadrul seriei. Indicatorii simpli sunt ***amplitudinea***, ***abaterea interquartilă*** şi ***abaterile individuale***. Toți indicatori pot fi exprimați în mărimi absolute (adică în unitatea de măsură a caracteristicii analizate) sau în mărimi relative, calculate în raport cu media sau mediana.

AMPLITUDINEA

Amplitudinea (A),[1] se obține prin diferența dintre valoarea cea mai mare şi cea mai mică a caracteristicii respective.

Amplitudinea absolută: $$A = x_{\max} - x_{\min} \qquad (5.1)$$

Amplitudinea relativă: $$A_r = \frac{x_{\max} - x_{\min}}{m} \qquad (5.1')$$

Acest indicator este cel mai simplu de calculat dar şi cel mai dezavantajos, deoarece ține seama doar de două valori, cele extreme, fără a oferii informații despre termenii din interiorul seriei.

Iată două serii statistice (de exemplu: notele obținute de elevi unei clase la două discipline diferite) care au aceeaşi amplitudine:

prima serie:	2 3 4 4 4 5 5 6 6 6 6 7 7 8 8 8 9 9 10
a doua serie:	2 5 5 5 5 5 6 6 6 6 6 6 6 6 7 7 7 7 10

În ambele cazuri amplitudinea va fi 8 ($A = x_{max} – x_{min} = 10 – 2 = 8$), însă prima serie prezintă o variație reală a notelor, pe când în cea de-a doua valorile extreme pot fi considerate excepții (atipice), nivelul redus al variație nefiind reflectat deloc în valoarea amplitudinii.

Din aceste motive, utilizarea amplitudinii în vederea caracterizării omogenității/eterogenității unei serii statistice trebuie făcută cu rezerve, doar atunci când valorile extreme nu se abat foarte mult de la ceilalți termeni ai seriei.

ABATEREA INTERQUARTILĂ

Abaterea interquartilă (I) sau **abaterea quartilă**, se obține prin diferența dintre quartila cea mai mare şi cea mai mică a caracteristicii respective[2]. După cum am aflat în capitolul anterior, quartilele sunt în număr de trei (notate Q_1, Q_2, Q_3); ele împart seria statistică în patru părți egale (vezi cap. 4.2.). Reamintim că Q_2 este de fapt mediana seriei.

---

[1] În engleză: *Range*.

[2] Similar pot fi definite abaterile interdecile sau intercentile.

Abaterea interquartilă absolută: $I = Q_3 - Q_1$ (5.2)

Abaterea interquartilă relativă: $I_r = \frac{Q_3 - Q_1}{Q_2}$ (5.2')

Prin utilizarea acestui indicator sunt eliminate valorile extreme, mai precis, valorile situate în primul sfert (între x_{min} şi Q_1) şi ultimul sfert (între Q_3 şi x_{max}) al seriei, reducându-se astfel influența acestora. Abaterea interquartilă este preferată în locul amplitudinii atunci când valorile extreme din cadrul seriei sunt atipice, adică se abat prea mult de la ceilalți termeni ai seriei. Acest indicator este reprezentat grafic cu ajutorul diagramelor de tip *Boxplot* (vezi capitolul 4.5.).

Reluând exemplul de mai sus, pentru a doua serie statistică abaterea interquartilă este $I = Q_3 - Q_1 = 7 - 5 = 2$, ceea ce reflectă mult mai bine lipsa de variație a valorilor seriei.

$\underline{2}$	5	5	5	$\underline{5}$	5	6	6	6	$\underline{6}$	6	6	6	6	$\underline{7}$	7	7	7	$\underline{10}$
x_{min}				Q_1					$Q_2 = M_e$					Q_3				x_{max}

Cu toate acestea, nici în acest caz nu avem informații despre ce se întâmplă între cele două quartile extreme, mai mult, apare dezavantajul eliminării a jumătate din termenii seriei (din acest motiv, uneori calculăm abaterea interdecilă, care elimină o cincime dintre valori, sau chiar abaterea intercentilă, aceasta eliminând doar a cincizecia parte dintre valori).

Toate aceste dezavantaje induse de amplitudine şi de abaterea interquartilă pot fi eliminate dacă se calculează abaterile (diferențele) nu doar dintre două valori, ci între toate valorile seriei respective. Se obține astfel un indicator cunoscut sub numele de ***indicele lui Gini***[3], mai puțin folosit de către psihologi, sociologi sau pedagogi. Mai cunoscute sunt acele abateri calculate pentru toate valorile caracteristicii prin raportare la o valoare fixă, de obicei media sau mediana.

ABATERILE INDIVIDUALE

Abaterile (deviațiile) individuale (d_i), mai precis **abaterile individuale de la medie**[4], se obțin prin diferența dintre fiecare valoare şi media aritmetică a caracteristicii respective. La fel pot fi calculate abaterile individuale de la mediană sau de la oricare altă valoare din cadrul seriei.

Conform proprietăților mediei (vezi capitolul 4.1.) suma acestor abateri individuale este întotdeauna egală cu zero.

[3] Indicele lui Gini (după numele statisticianului italian Corado Gini) este definit ca: media aritmetică a diferențelor dintre toate perechile de valorii, diferențe luate în valoare absolută/în modul (pentru formule vezi T. Rotariu *et. al.*, 1999, p. 52).

[4] În practica statistică cele mai dese abateri individuale sunt calculate în raport cu media aritmetică, din acest motiv de cele mai multe ori, pentru a simplifica, vom folosi termenul de abatere individuală în locul celui de abatere individuală de la medie.

Abaterile individuale absolute: $d_i = x_i - m$ (5.3)

Abaterile individuale relative: $d_{ir} = \frac{x_i - m}{m}$ (5.3')

Abaterile individuale ne oferă informații doar despre poziția unuia sau altuia dintre subiecți în raport cu media seriei, fără însă a surprinde în mod sintetic gradul de variație al caracteristicii. Pentru aceasta trebuie considerate toate abaterile individuale ale valorilor caracteristicii de la media lor, lucru posibil de realizat doar cu ajutorul indicatorilor sintetici ai variației.

5.2. INDICATORI SINTETICI AI VARIAȚIEI

Aceşti indicatori au la bază calcularea valorii medii a tuturor abaterilor individuale ale variantelor de la media lor (se poate lua ca reper şi mediana seriei sau oricare altă valoare a seriei!). Se realizează astfel o sintetizare a variației unei caracteristici printr-o singură expresie numerică.

Indicatorii sintetici sunt ***abaterea medie liniară***, ***dispersia***, ***abaterea medie pătratică*** şi ***coeficientul de variație***. Vom prezenta formulele pentru seriile simple şi pentru seriile (distribuțiile) de frecvențe.

ABATEREA MEDIE LINIARĂ

Abaterea (deviația) medie liniară (d) sau pur şi simplu **abaterea medie**,[5] se calculează ca o media aritmetică a tuturor abaterilor individuale, luate în valoare absolută (fără a lua în considerare semnul – sau +).

Abaterea medie în cazul seriilor simple: $d = \frac{\sum |x_i - m|}{n}$ (5.4)

Abaterea medie în cazul seriilor de frecvențe: $d = \frac{\sum |x_i - m| \cdot f_i}{\sum f_i}$ (5.4')

Prin luarea în considerare a valorilor absolute se elimină, de fapt, acel inconvenient generat de proprietatea mediei aritmetice prin care suma abaterilor individuale este întotdeauna egală cu zero, adică $\sum x_i - m = 0$.

Abaterea medie ne arată cu cât se abate în medie fiecare valoare de la nivelul mediu şi se exprimă în unitatea de măsură a caracteristicii studiate. Dezavantajul acestui indicator constă în faptul că el acordă aceeaşi importanță tuturor abaterilor

[5] Şi de data aceasta, pentru simplificare, atunci când folosim termenul de abatere medie ne referim la abaterea medie de la medie. Se poate calcula abaterea medie de la mediană sau de la oricare altă valoare a seriei.

individuale, fără să țină seama de abaterile individuale mai mari care, în valoare absolută, influențează în mai mare măsură gradul de variație.

DISPERSIA

Dispersia (s^2 sau σ^2**)**[6] sau **varianța**,[7] se calculează ca o medie aritmetică a pătratelor abaterilor individuale ale tuturor valorilor față de media lor.

Dispersia în cazul seriilor simple:

$$s^2 = \frac{\sum (x_i - m)^2}{n} \qquad (5.5)$$

Dispersia în cazul seriilor de frecvențe:

$$s^2 = \frac{\sum (x_i - m)^2 \cdot f_i}{\sum f_i} \qquad (5.5')$$

Estimarea dispersiei unei populații, calculată pe baza unui eșantion [8]:

$$s^2 = \frac{\sum (x_i - m)^2}{n-1} \qquad (5.6)$$

$$s^2 = \frac{\sum (x_i - m)^2 \cdot f_i}{(\sum f_i) - 1} \qquad (5.6')$$

Sunt autori care susțin că termenul de dispersie ar trebui evitat deoarece el „*este unul generic, fiind utilizat pentru toți indicatorii din categoria celor care reflectă împrăştierea valorilor*" (Rotariu *et.al.*, 1999, p. 42). Pe de altă parte, varianța reprezintă „*indicatorul sintetic de bază al dispersiei*" (Luduşan *et.al.*, 1997, p. 277) sau „*indicatorul statistic cel mai utilizat pentru aprecierea împrăştierii datelor*" (Clocotici & Stan, 2000, p. 68).

Dincolo de aceste opinii divergente, suntem de părere că el nu trebuie neglijat, oferindu-ne date despre gradul de omogenitate/eterogenitate al caracteristicii vizate; utilitatea lui o vom vedea la calculul următorului indicator şi în capitolele de statistică inferențială.

---

[6] Se foloseşte s^2 când facem referire la un eşantion şi σ^2 (sigma la pătrat) când calculăm abaterea standard pentru întreaga populație. Aceeaşi semnificație o au şi notațiile pentru abaterea standard: s şi σ.

[7] În engleză: *variance.*

[8] Programele statistice pentru prelucrarea informatizată a datelor (SPSS, Excel etc.) folosesc pentru calculul dispersie şi abaterii standard formule ce au la numitor *n*-1. Este o corecție generată de considerente teoretice - vezi caseta 5.1. Prin aceste formule se obțin estimări ale celor doi indicatori la nivelul întregii populații statistice, în condițiile în care valorile la care ne raportăm aparțin unui eșantion extras din populația respectivă.

ABATEREA STANDARD

Abaterea standard[9] **(s** sau **σ)**, numită şi **abaterea medie pătratică** sau **abaterea tip**,[10] reprezintă rădăcina pătrată din valoarea dispersiei.

Abatere medie pătratică în cazul seriilor simple:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - m)^2}{n}} \qquad (5.7)$$

Abaterea medie pătratică în cazul seriilor de frecvențe:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - m)^2 \cdot f_i}{\sum f_i}} \qquad (5.7')$$

Estimarea abaterii standard a unei populații, calculată pe baza unui eşantion:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - m)^2}{n-1}} \qquad (5.8)$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - m)^2 \cdot f_i}{(\sum f_i) - 1}} \qquad (5.8')$$

Proprietățile abaterii standard:

- dacă la toate valorile seriei statistice se adaugă (scade) o constantă *c*, abaterea standard nu se modifică: dacă $y_i = x_i + c$ sau $y_i = x_i - c$, atunci $s_y = s_x$
- dacă toate valorile seriei statistice se înmulțesc/divid cu o constantă *c*, atunci şi abaterea standard se va multiplica/divide cu aceeaşi valoare *c*: dacă $y_i = c \cdot x_i$, atunci $s_y = c \cdot s_x$
- abaterea standard față de medie este mai mică decât abaterea standard față de oricare altă valoare (mediană etc.) a distribuției.

Mult mai des folosită în analiza seriilor statistice, abaterea medie pătratică are acelaşi avantaj ca şi abaterea medie liniară, şi anume, se exprimă în aceeaşi unitate de măsură ca şi datele inițiale pe care le studiem. De exemplu, dacă studiul se bazează pe notele unui colectiv de elevi, abaterea tip se exprimă tot în note,

[9] În engleză: *standard deviation (SD).*
Abaterea standard se referă doar la abaterea medie pătratică față de medie. Putem calcula şi abaterea medie pătratică față de mediană, prin înlocuirea mediei cu mediana.

[10] În franceză: *écart type.*

„permiţând să se analizeze mai corect gradul de variabilitate al grupului" (Radu *et.al.*, 1993, p.72).

Asemănător dispersiei, o valoarea scăzută a abaterii standard reflectă o serie statistică omogenă; în caz contrar vorbim de eterogenitatea datelor. Mai mult, pe graficul distribuţiei acest indice marchează punctele de inflexiune ale curbei.

Totuşi, atunci când dorim să comparăm serii statistice cu unităţi de măsură diferite, ultimii doi indicatori nu ne mai sunt de folos. Vom folosi un alt indicator: coeficientul de variaţie.

COEFICIENTUL DE VARIAŢIE (DE VARIABILITATE)

Coeficientul de variaţie (V) reprezintă raportul dintre abaterea medie pătratică şi media colectivităţii studiate. Se foloseşte atunci când dorim să comparăm gradul de împrăştiere al unor serii statistice exprimate în unităţi de măsură diferite (de exemplu: înălţimile a două eşantioane de subiecţi, exprimate în centimetrii, respectiv în inch). De asemenea, utilizăm acest indicator şi când seriile statistice au aceeaşi unitate de măsură, dar nivelul general al valorilor caracteristicii studiate este total diferit (de exemplu: înălţimile unor copii de la grădiniţă şi cele ale unor elevi de liceu, exprimate în centimetri).

Coeficientul de variaţie: $$V = \frac{s}{m} \cdot 100 \qquad (5.9)$$

Acest indicator se exprimă în procente (se poate elimina înmulţirea cu 100; vom obţine valori între 0 şi 1) şi ne arată gradul de omogenitate/eterogenitate al colectivităţii statistice studiate, astfel: cu cât valoarea coeficientului de variaţie este mai aproape de zero, cu atât variaţia este mai mică, deci colectivitatea este mai omogenă.

Dacă coeficientul de variaţie este cuprins între 0 şi 15%, înseamnă că împrăştierea datelor este foarte mică, iar media este reprezentativă, deoarece eşantionul măsurat este omogen. Dacă valoarea lui este între 15 şi 30%, împrăştierea datelor este mijlocie, media fiind încă suficient de reprezentativă. Limita maximă admisă pentru ca un eşantion să fie considerat omogen iar media să fie reprezentativă pentru colectivitatea respectivă este de 35% (Novak, 1995).

Nici acest ultim indicator nu este lipsit de contraindicaţii! Cel puţin două atenţionări trebuie făcute:

- formula coeficientului de variaţie este aplicabilă doar în cazul variabilelor măsurate pe scale de rapoarte, cu origine zero naturală (rar întâlnite în psihologie şi pedagogie);
- nu oricare două caracteristici pot fi comparate cu ajutorul coeficientului de variaţie (de exemplu: *este inutil să comparăm un eşantion după salariul membrilor cu alt eşantion în care avem în vedere numărul de la pantofi!* – cf. Rotariu *et.al.*, 1999, p. 59).

5.3. INDICATORI AI FORMEI DISTRIBUŢIEI

Gradul de împrăştiere a valorilor unor serii statistice determină şi forme diferite ale reprezentărilor grafice ataşate acestor distribuţii statistice. Pentru a reflecta forma

unei distribuții, mai ales pentru a face comparații între două sau mai multe serii, ne folosim de o altă categorie de indicatori, numiți **indicatori ai formei**. Cei doi indicatori folosiți în statistica socială sunt: oblicitatea şi boltirea.

INDICATORUL OBLICITĂȚII (DE ASIMETRIE)

Oblicitatea[11] a fost propusă de către Pearson pentru aprecierea gradului de simterie/asimetrie a unei serii statistice. Se calculează cu una din formulele:

Oblicitatea:

$$O = \frac{3 \cdot (m - M_e)}{s} \qquad (5.10)$$

sau

$$O = \frac{m - M_o}{s} \qquad (5.10')$$

sau

$$O = \frac{\sum (x_i - m)^3}{ns^3} \qquad (5.10'')$$

Prin ridicarea abaterilor individuale la puterea a treia (formula 5.10'') se acordă o mai mare importanță valorilor extreme. Putem analiza astfel gradul de asimetrie al distribuției, altfel spus, tendința valorilor de a se grupa spre una din cele două extreme.

În cazul distribuțiilor simetrice, deoarece media şi modul sunt identice, oblicitatea va fi 0. În cazul curbelor de distribuție asimetrice, alungite spre dreapta sau spre stânga, oblicitatea va avea o valoarea negativă, respectiv pozitivă (vezi cap. 6.2.).

INDICATORUL BOLTIRII (DE EXCES, DE APLATIZARE)

Boltirea[12] exprimă înălțimea „cocoaşei" curbei de distribuție, comparativ cu cea normală. Ne arată măsura în care o distribuție este mai plată sau mai boltită.

Boltirea:

$$B = \frac{\sum (x_i - m)^4}{ns^4} - 3 \qquad (5.11)$$

Pentru valori pozitive ale acestui indicator spunem că avem o distribuție „leptokurtică" (cu cocoaşă înaltă). În celălalt sens, distribuția va fi „platikurtică" (cu cocoaşă aplatizată) – vezi figura 5.1. Valori apropiate de 0 indică o distribuție „mezokurtică"

[11] În engleză: *skewness.*
[12] În engleză: *kurtosis* (=cocoaşă).

Sunt considerate distribuții relativ normale cazurile în care aceşti indicatori nu depăşesc ±1,96.

5.4. UTILIZAREA SPSS PENTRU CALCULAREA INDICATORILOR VARAȚIEI ŞI AI FORMEI

Şi de această dată dispunem de mai multe posibilități pentru a calcula indicatorii variației sau pe cei ai formei unei serii statistice.

Ca şi în capitolele anterioare, prezentăm pentru început soluția parcurgerii următoarelor comenzi:

„Analyze" – „Descriptive Statistics" – „Frequencies…"

După ce, în fereastra de dialog pentru calculul frecvențelor (vezi figura 2.1.), selectăm variabila sau variabilele dorite, apăsăm butonul „*Statistics…*" şi vom pătrunde într-o nouă fereastră de opțiuni (figura 5.1).

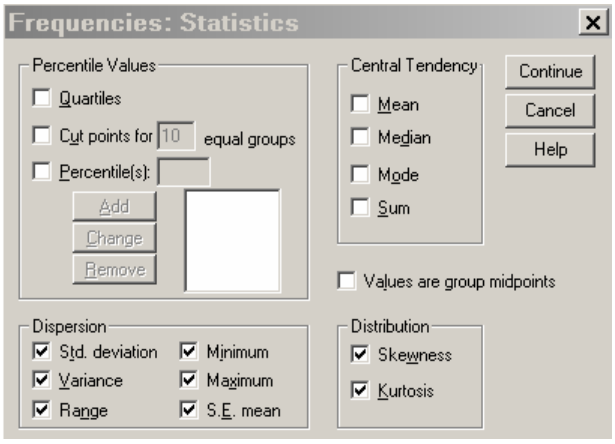


Figura 5.1. *Fereastră de opțiuni pentru calculul unor indicatori statistici*

La rubrica „*Dispersion*" putem opta pentru calculul abaterii standard (*Std. deviation*), a variației, a amplitudinii (*Range*), a valorilor minime şi maxime şi a erorii standard a mediei (*S.E. mean*).

La rubrica „*Distribution*" se optează pentru calcularea oblicității (*Skewness*) sau boltirii (*Kurtosis*).

6.

DISTRIBUȚIILE STATISTICE

6.1. Distribuția normală
6.2. Distribuții simetrice şi asimetrice
6.3. Distribuții unimodale şi bimodale
6.4. Valori normate (scoruri z)
6.5. Distribuția normală standardizată

După cum am arătat în capitolele anterioare (capitolul 3), prin asocierea variantelor (valorilor) unei variabile statistice cu frecvențele (absolute sau relative) cu care acestea apar se obține o **DISTRIBUȚIE STATISTICĂ**. Pentru exprimarea sintetică a informațiilor conținute de aceste şiruri de date putem calcula o mulțime de indicatori statistici, astfel încât, printr-o simplă analiză a lor să putem spune dacă distribuțiile statistice sunt simetrice sau asimetrice, unimodale sau multimodale, aplatizate sau înalte.

6.1. DISTRIBUȚIA NORMALĂ

Cunoscută şi sub denumirea de **curba (clopotul) lui Gauss**, este o distribuție simetrică, spre care tind toate şirurile de date obținute în practica statistică şi care se caracterizează prin aceea că valorile centrale sunt cât mai apropiate, iar de o parte şi de alta a lor avem un număr aproximativ egal de valori. Într-o distribuție perfect normală[1] media, mediana şi modul sunt identice, iar celelalte valori sunt dispuse perfect simetric de o parte şi de alta a acelei valori centrale.

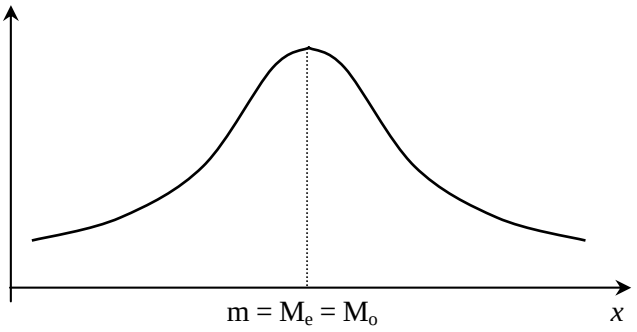


Figura 6.1 *Curba distribuției normale*

[1] Distribuția perfect normală este o distribuție teoretică unimodală, simetrică şi continuă.

Matematicianul K.F. Gauss a constatat următorul aspect: cu cât obținem mai multe valori ale caracteristicii respective, cu atât curba distribuție tinde spre cea perfect normală (sau teoretică). De altfel, acest tip de curbă este considerat de cele mai multe ori ca un reper, normalitatea unei distribuții verificându-se față de această curbă perfect simetrică sau, altfel spus, distribuția normală reprezintă o bună aproximație pentru distribuțiile multor variabile întâlnite în aplicațiile statistice curente.

Caracteristicile curbei normale şi frecvența cu care se face apel la aceasta în studiile statistice determină adesea interpretări greşite. Atragem atenția că distribuțiile reale pe care le descoperă psihologii în studiile lor nu au niciodată parametrii unei curbe normale perfecte. Acest lucru este practic imposibil dacă ne gândim că o curbă normală are limitele deschise, mergând spre infinit, în timp ce distribuțiile reale sunt finite (Popa, 2004).

6.2. DISTRIBUȚII SIMETRICE ŞI ASIMETRICE

În analiza fenomenele psihosociale distribuțiile devin simetrice (vezi distribuția normală), de cele mai multe ori, doar dacă cercetătorul analizează un număr suficient de mare de cazuri, astfel încât indicatorii tendințelor centrale să coincidă, iar de o parte şi de alta a lor să avem un număr aproximativ egal de valori.

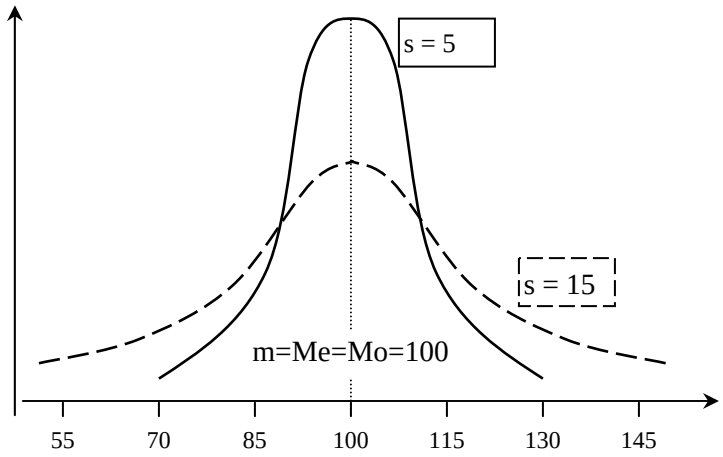


Figura 6.2. *Curbe de distribuție simetrice*

În foarte multe situații, însă, variantele cu cele mai mari frecvențe (valorile sau intervalele modale) nu coincid cu celelalte valori centrale (media sau mediana) înregistrându-se o polarizarea spre dreapta sau spre stânga a acestora. Pot apărea următoarele două situații:

- $m > M_e > M_o$ – spunem că distribuția prezintă o asimetrie de stânga sau pozitivă;
- $m < M_e < M_o$ – spunem că distribuția prezintă o asimetrie de dreapta sau negativă (figura 6.3).

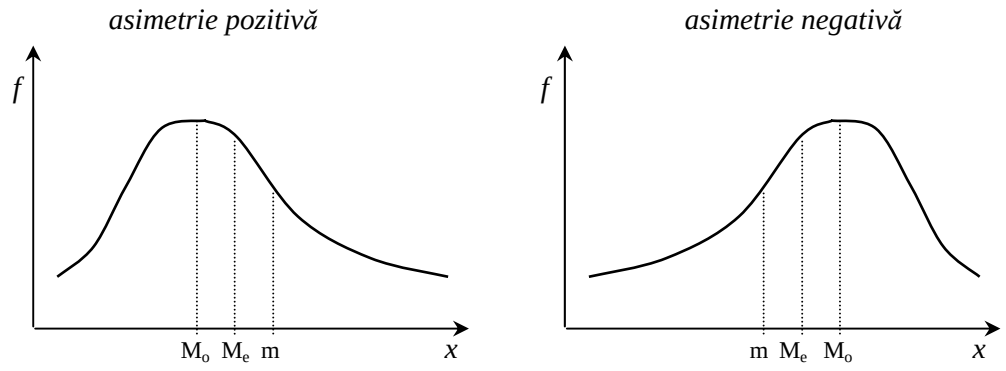


Figura 6.3. *Curbe de distribuție asimetrice*

Reamintim că acest grad de asimetrie ne este dat şi de un indicator al formei distribuției şi anume, oblicitatea (vezi 5.3.). Acesta, prin valorile pozitive sau negative pe care le ia, ilustrează asimetria pozitivă sau negativă.

O asimetrie accentuată spre stânga sau spre dreapta determină apariția unor tipuri particulare de distribuții, cunoscute cu numele de distribuții în formă de „*i*" şi în formă de „*j*" (figura 6.4.). De exemplu, erorile pe parcursul unui proces de formare a unei deprinderi sau timpul de execuție al unei acțiuni în procesul exercițiului vor înregistra valori constant descrescătoare, astfel încât, reprezentarea grafică a variației lor va avea forma literei „*i*" (Radu *et.al.*, 1993).

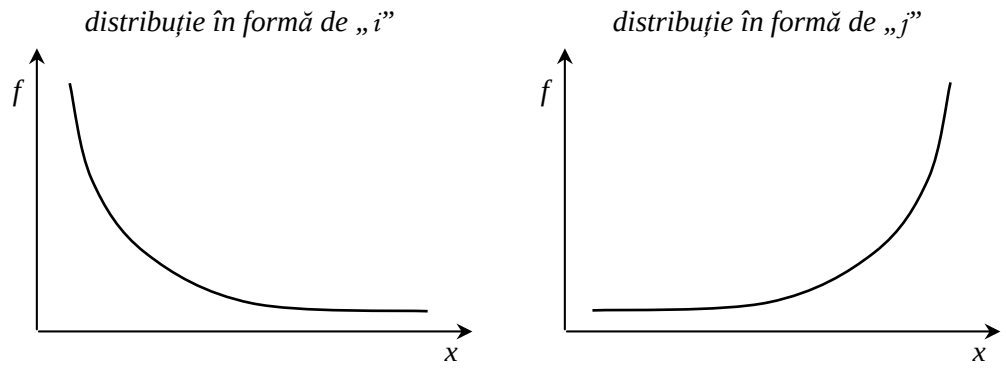


Figura 6.4. *Curbe de distribuție în formă de „i" şi „j"*

6.3. DISTRIBUȚII UNIMODALE ŞI BIMODALE

În unele serii statistice media îşi pierde reprezentativitatea deoarece colectivitatea are tendința de a se grupa în două (sau mai multe) grupe distincte. De data aceasta modul este indicatorul de poziție cel mai relevant. Din acest motiv, vom spune că avem de-a face cu o **DISTRIBUȚIE BIMODALĂ** (uneori chiar **multimodală**).

La rândul lor, distribuțiile bimodale pot fi simetrice sau asimetrice, negative sau pozitive (figura 6.5.)

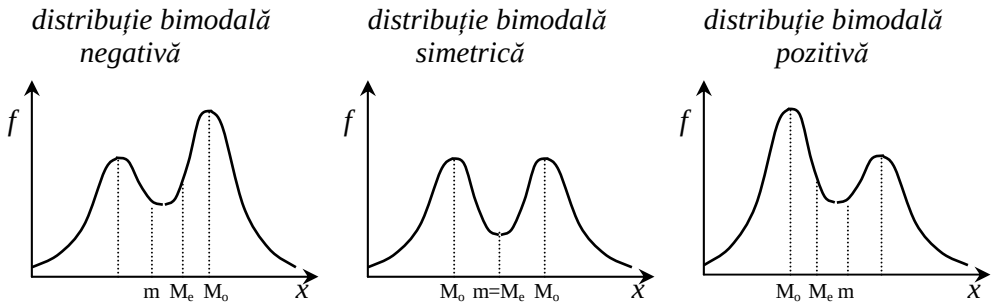


Figura 6.5. *Curbe de distribuție bimodale*

Încheiem această prezentare a tipurilor de distribuții statistice cu precizarea că în cazul curbelor simetrice se recomandă determinarea mediei şi a abaterii standard, în timp ce pentru seriile statistice asimetrice sunt preferate valorile medianei şi oblicității. În cazul curbelor de distribuție în formă de „*i*”, a celor în formă de „*j*” şi a celor bimodale este bine să ne mulțumim cu un grafic şi să determinăm modul, respectiv frecvențele (Radu *et.al.*, 1993).

6.4. VALORILE NORMATE (STANDARDIZATE) – SCORURI Z

De foarte multe ori suntem puși în situația de a compara valori ale unor caracteristici psihologice despre care nu cunoaştem mare lucru. De exemplu, scorul de 17 puncte obținut de un subiect pe scala de introversie/extraversie nu ne îndreptățeşte să afirmăm că este un scor mare sau mic, şi nici că este mai bun sau mai rău decât cel de 9 puncte obținut, de acelaşi subiect, pe scala de stabilitate/instabilitate.

În situația în care nu cunoaştem semnificația datelor colectate în formă brută putem recurge la transformarea acestora din cote brute în **valori normate (standardizate)**, transformare ce se bazează pe proprietățile mediei şi abaterii standard, în cazul unei distribuții normale.

Scorul normat z (numit şi **cota z** sau **scor z**) exprimă semnificația unei anumite valori dintr-o distribuție prin raportare la parametrii distribuției (medie şi abatere standard). Altfel spus, aceasta măsoară distanța dintre o anumită valoare şi media distribuției, în abateri standard. Formula de calcul este:

$$z = \frac{x - m}{s} \qquad (6.1)$$

unde x reprezintă oricare dintre valorile distribuției,
m şi s reprezintă media, respectiv abaterea standard.

Scorul z se numeşte şi „**scor standardizat z**” (**notă standardizată z**). Aceasta pentru că poate fi utilizat pentru a compara valori care provin din distribuții diferite, indiferent de unitatea de măsură a fiecăreia.

Exemplu (apud Sava, 2004a): Un subiect a obținut 43 de răspunsuri corecte la un test de acuitate vizuală (TAV) şi 18 puncte la un test de atenție concentrată (TAC). Dacă transformăm în cote z cele 43 de puncte obținute la TAV, vom obține valoarea -1,71 (ştiind că m = 55, s = 7). Similar, dacă vom transforma în cote z rezultatul obținut la TAC, vom obține -0,96 (m = 21, s = 3,11). Pe baza acestor transformări putem afirma că, deşi ambele rezultate sunt sub medie, performanța la TAC este mai bună decât cea obținută la TAV.

Utilizând proprietățile de transformare a formulei de definiție a scorului z, putem calcula o anumită valoare atunci când cunoaştem valoarea lui z şi parametrii distribuției, astfel:

$$x = z \cdot s + m \qquad (6.2)$$

Proprietățile scorurilor z

1. *Media unei distribuții z este întotdeauna egală cu 0.*

   Pentru a explica această afirmație facem apel la una dintre proprietățile mediei, şi anume: scăderea unei constante la fiecare valoare determină scăderea mediei cu acea valoare (vezi 4.1.). Formula de calcul pentru z implică scăderea unei constante din fiecare valoare a distribuției. Aceasta înseamnă că şi media noii distribuții (z) se va reduce cu constanta respectivă. Dar această constantă este însăşi media distribuției originale, ceea ce înseamnă că distribuția z va avea media egală cu zero, ca rezultat al diminuării mediei cu ea însăşi.

2. *Abaterea standard a unei distribuții z este întotdeauna 1*.

   Acest fapt decurge prin efectul cumulat al proprietăților abaterii standard (vezi 5.2.). Prima proprietate afirmă că în cazul scăderii unei constante (în cazul scorurilor z, media) din valorile unei distribuții, abaterea standard a acesteia nu se modifică. A doua proprietate afirmă că în cazul împărțirii valorilor unei distribuții la o constantă, noua abatere standard este rezultatul raportului dintre vechea abatere standard şi constantă. Dar constanta de care vorbim este, în cazul distribuției z, chiar abaterea standard. Ca urmare, noua abatere standard este un raport dintre două valori identice al cărui rezultat, evident, este 1. (Popa, 1996)

Alte tipuri de scoruri standardizate

Cotele z prezintă doua avantaje importante: permit compararea valorilor unei distribuții, şi a valorilor provenite din distribuții diferite, ca urmare a faptului că se exprimă în abateri standard de la medie. Totuşi se impune o anumită precauție în comparația pe baza scorurilor z atunci când distribuțiile au forme diferite şi, mai ales, asimetrii opuse.

Notele z au, însă, şi unele dezavantaje: se exprimă prin numere mici, cu zecimale, (greu de manipulat intuitiv) şi, în plus, pot lua valori negative. Aceste dezavantaje pot fi uşor înlăturate printr-un artificiu de calcul care să conducă la note standardizate convenabile (ce corespund anumitor nevoi specifice). Mai jos sunt descrise câteva tipuri de note standard calculate pe baza notelor z.

Cote T (Thurstone) – media unei distribuții T este întotdeauna egală cu 50 iar abaterea standard cu 10.

$$T = 50 + 10 * z \qquad T = 50 + 10 * \frac{x - m}{s} \qquad (6.3)$$

Cote H (Hull) – media unei distribuții H este întotdeauna egală cu 50 iar abaterea standard cu 14.

$$H = 50 + 14 * z \qquad H = 50 + 14 * \frac{x - m}{s} \qquad (6.4)$$

Cote IQ (Binet) – media unei distribuții IQ de acest tip este întotdeauna egală cu 100 iar abaterea standard cu 16.

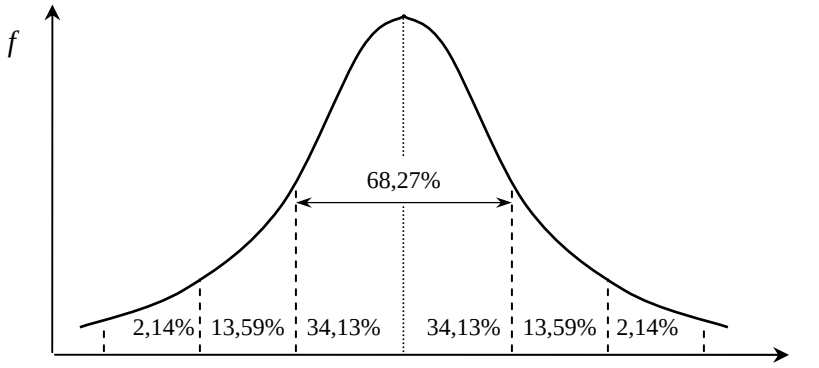
$$IQ = 100 + 16 * z \qquad IQ = 100 + 16 * \frac{x - m}{s} \qquad (6.5)$$

Cote IQ (Wechsler) – media unei distribuții IQ de acest tip este întotdeauna egală cu 100 iar abaterea standard cu 15.

$$IQ = 100 + 15 * z \qquad IQ = 100 + 15 * \frac{x - m}{s} \qquad (6.6)$$

6.5. DISTRIBUȚIA NORMALĂ STANDARDIZATĂ

Distribuția normală în care valorile sunt exprimate în scoruri z se numeşte **CURBĂ NORMALĂ STANDARDIZATĂ**. Ea are toate proprietățile enunțate mai sus, având însă şi parametrii oricărei distribuții z: m=0 şi s=1. Valoarea 0 pentru medie a fost aleasă convențional pentru că astfel distribuția este simetrică în jurul lui 0.



cote z	-3	-2	-1	0	+1	+2	+3
cote T	20	30	40	50	60	70	80

Figura 6.6. *Curba distribuției normale*

Curba normală standardizată are câteva caracteristici care sunt figurate în imaginea de mai sus şi pe care este important să le reținem:

- 34,13% dintre scorurile distribuției normale se află între medie şi o abatere standard deasupra mediei (z = +1). La fel pentru z = –1.
- Între –1z şi +1z se află aproximativ 68% dintre valorile distribuției.
- Aproximativ 96% dintre scoruri se află între –2z şi +2z.

Mai mult, pe baza distribuției scorurilor z pe o curbă normală standardizată putem preciza:

- procentajul de valori care se află sub/peste o anumită cotă z;
- procentajul de valori care se află între anumite cote z; ori între medie şi o cotă z
- cota z corespunzătoare unui anumit procentaj de valori.

Pentru aceasta, utilizăm un tabel special în care sunt trecute ariile determinate de curba distribuției normale ce corespund distanței dintre medie şi z abateri standard de la medie. Aceste cifre exprimă, sub formă de probabilități, frecvențele valorilor de sub curba normală z (Anexa 1).

Aria de sub curba normală văzută ca probabilitate

Valorile reprezentate pe curba normală nu constituie valori reale, rezultate în urma unui proces de măsurare. Ele reprezintă valori ipotetice, distribuite astfel pe

baza unui model matematic (legea numerelor mari). Nimic nu ne împiedică să considerăm că valorile de sub curba normală sunt rezultatul unei ipotetice extrageri aleatoare. Pe măsură ce „extragem" mai multe valori, curba de distribuție a acestora ia o formă care se apropie de forma curbei normale. Extrăgând „la infinit" valori aleatoare, vom obține o distribuție normală perfectă, exprimabilă printr-o curbă normală perfectă.

Din cele spuse mai sus, rezultă faptul că valorile din zona centrală a curbei sunt mai „frecvente" (mai multe), pentru că apariția lor la o extragere aleatoare este mai „probabilă". În acelaşi timp, valorile „mai puțin probabile", apar mai rar, şi populează zone din ce în ce mai extreme ale distribuției (curbei).

Probabilitatea înseamnă „frecvența relativă a apariției unui eveniment". Subiectiv, se traduce prin „cât de siguri putem fi că acel eveniment apare".

Dacă probabilitatea reprezintă raportul dintre evenimentul favorabil şi toate evenimentele posibile, atunci valoarea ei variază între 0 şi 1. Ea poate fi exprimată şi în procente. De exemplu, probabilitatea de 0,05 corespunde unui procentaj de apariție de 5%

Utilizând simbolul p (de la „probabilitate"), spunem că dacă $p<0,05$ înseamnă că evenimentul are mai puțin de 5% şanse să apară, în condițiile unei distribuții corespunzătoare curbei normale.

Procentajul ariilor de sub curba normală poate fi citit, deci, şi ca probabilitatea a distribuției. De exemplu, probabilitatea de a avea un scor între medie şi z=+1 este de 0,3413, ceea ce înseamnă că pentru un scor z ales la întâmplare există 34,13 şanse dintr-o sută ca acesta să cadă în suprafața haşurată. (vezi figura 6.7. şi anexa 1)

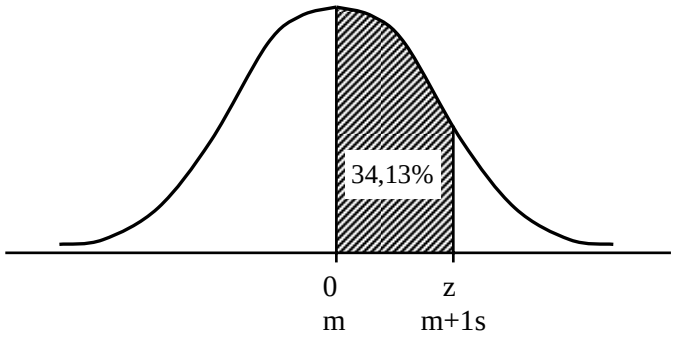


Figura 6.7. *Probabilitatea de a avea un scor între medie şi z=+1*

În acelaşi mod, pe baza proprietăților distribuției normale, vrem să identificăm valorile +z şi -z pentru care, într-o distribuție normală standardizată avem **95%**, respectiv **99%**, din valori. De aceste două repere, frecvent utilizate în statistica inferențială, se leagă probabilitățile de 5%, respectiv 1%.

Vom identifica aceste două repere cu ajutorul anexei 1:

- pentru **z=1,96** aria de sub curba normală delimitată de medie şi +z este de 0,4750; adică 47,5% din valorile z sunt cuprinse între 0 şi 1,96 şi tot atâtea între -1,96 şi 0;
- pentru **z=2,58** aria de sub curba normală delimitată de medie şi +z este de aprox. 0,4950; adică 49,5% din valorile z sunt cuprinse între 0 şi 2,58 şi tot atâtea între -2,58 şi 0.

Altfel spus: într-o distribuție normală standardizată, 95% dintre valorile z sunt cuprinse între -1,96 şi 1,96; de asemenea, avem 99% dintre valorile z cuprinse între -2,58 şi 2,58. Putem scrie aceste relații sub forma:

$$-1{,}96 < z < 1{,}96$$

ne folosim de formula 6.1. pentru a obține:

$$-1{,}96 < (x - m) / s < 1{,}96$$

$$(m - 1{,}96s) < x < (m + 1{,}96s) \qquad (6.7)$$

Deci, pentru o distribuție normală a unei variabile oarecare (nestandardizată) concluziile de mai sus devin (vezi figura 6.8.):

- avem 95% din valorile *x* cuprinse în intervalul [m-1,96s; m+1,96s];
- avem 99% din valorile *x* cuprinse în intervalul [m-2,58s; m+2,58s].

Cu alte cuvinte, există 5% şanse ca o valoare *x* luată la întâmplare să fie în afara intervalului [m-1,96s; m+1,96s], după cum există o şansă din 100 ca |*x*| să fie mai mare ca m+2,58s.

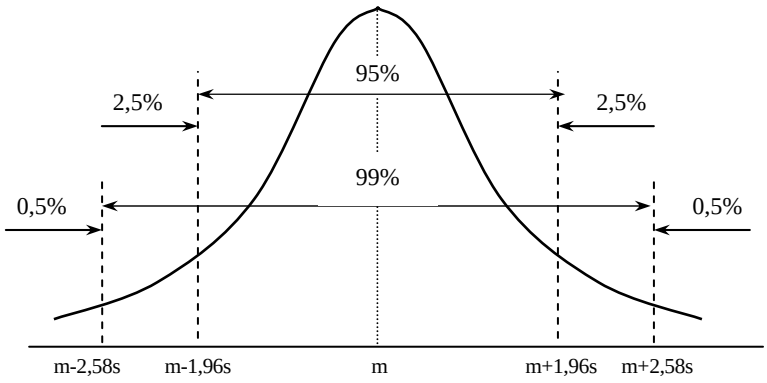


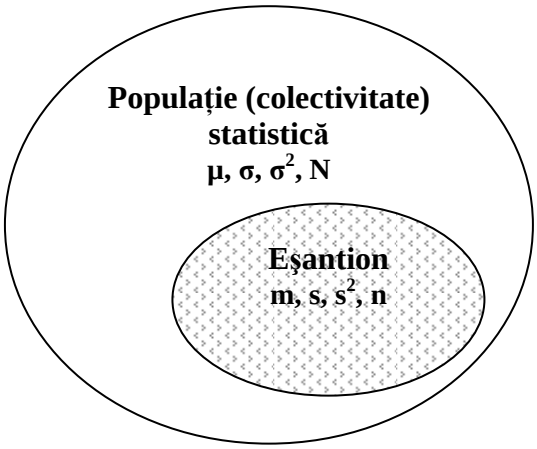
Figura 6.8. *Proprietățile distribuției normale*

7. INFERENȚA STATISTICĂ

7.1. Delimitări conceptuale
7.2. Probleme de estimare
7.3. Testarea ipotezelor
7.4. Testele parametrice t şi z.
  7.4.1. Testele t şi z pentru un eşantion.
  7.4.2. Testele t şi z pentru două eşantioane independente
  7.4.3. Testele t şi z pentru două eşantioane dependente
7.5. Utilizarea SPSS pentru aplicarea testului t

7.1. DELIMITĂRI CONCEPTUALE

Datele obținute în cursul unei experiențe, a unei observații sistematice sau anchete, constituie un *eşantion* extras dintr-o colectivitate mai largă sau *populație*. Pe de altă parte, statistica descriptivă, reduce datele brute la câteva valori caracteristice: frecvențe absolute sau relative, medii, abateri standard etc. Reamintim simbolurile pentru aceşti parametrii, în cele două situații: μ, σ, σ^2 – în cazul întregii colectivități statistice; m, s, s^2 – când ne referim la un eşantion.



Se pune întrebarea în ce măsură, plecând de la indicatorii eşantionului cercetat, putem formula concluzii asupra populației? Cu alte cuvinte, se pune întrebarea: în ce măsură datele obținute sunt relevante pentru populație? Operația prin care facem extrapolarea concluziilor de la eşantion la populație se numeşte ***inferență statistică***.

Inferența statistică se bazează pe teoria probabilităților, permițând desprinderea unor concluzii cu caracter probabilist. În practică, orice rezultat discutat în termeni de valori semnificative statistic la un prag de .05 sau .01 a corespuns unui demers

specific statisticii inferențiale. Principalele demersuri pe care se bazează statistica inferențială sunt *estimarea parametrilor statistici* şi *testarea ipotezelor* (Sava, 2004a).

Eşantioane independente şi eşantioane perechi

În multe cazuri psihologul este pus în situația de a compara între ele mediile sau frecvențele obținute într-un experiment, punându-şi, în final, întrebarea dacă diferențele constatate între grupul de control şi cel experimental sunt semnificative sau nu.

Apar următoarele situații:

1. dacă cele două eşantioane sunt alese la întâmplare pe baza caracteristicilor lor naturale (de exemplu, două clase paralele) spunem că avem ***eşantioane independente***.
2. dacă cele două eşantioane sunt în relație unul cu celălalt spunem că avem ***eşantioane dependente*** (sau ***eşantioane perechi***). Uzual, există trei situații în care avem de a face cu eşantioane dependente:
   a. Perechile naturale: acestea nu sunt realizate de experimentator ci există în mod natural.
   b. Perechile artificiale: acestea sunt realizate de către experimentator pentru a egaliza cât mai mult grupele de subiecți.
   c. Măsurători repetate: reprezintă cazul cel mai des întâlnit, în special în terapie şi recuperare. Este vorba în această situație de un singur grup de subiecți care vor fi testați de două ori (înainte şi după introducerea variabilei independente).

7.2. PROBLEME DE ESTIMARE

Este unanim acceptat faptul că atunci când calculăm indicatori statistici pentru un eşantion facem acest lucru cu o anumită probabilitate. Altfel spus, nu reuşim să determinăm exact parametrii caracteristici ai întregii colectivități. Indicatorii statistici calculați pentru un eşantion reprezintă *estimări* ale parametrilor populației.

Deoarece nu putem determina cu exactitate valoarea acestor parametri, vom încerca să stabilim un interval – numit şi *interval de încredere* – în care se găseşte cu certitudine parametrul respectiv. Cu cât acest interval este mai mic, cu atât informația noastră asupra adevăratei valori în populație este mai precisă.

7.2.1. Semnificația unei medii

Notând cu **μ** valoarea medie calculată pentru întreaga populație şi cu **m** media la nivelul eşantionului, diferența (**μ - m**) reprezintă eroarea pe care noi o comitem atunci când în loc să cercetăm toți cei **N** indivizi, prelevăm datele numai de la o subpopulație oarecare de **n** indivizi. De cele mai multe ori această eroare este diferită de 0, motiv pentru care devine necesară evaluarea ei. Însă, prin altă metodă decât făcând diferența (**μ - m**), deoarece întotdeauna media populației ne este necunoscută (dacă am cunoaşte valoarea lui **μ** nu s-ar mai pune problema estimării)

Semnificația unei valori medii depinde de doi parametrii:

- volumul eşantionului (n) pe care se calculează media şi

- abaterea standard (σ) calculată la nivelul întregii populații.

Cu cât volumul eşantionului este mai mare iar dispersia populației mai mică, cu atât media calculată la nivelul eşantionului devine mai reprezentativă pentru întreaga colectivitate (Radu *et.al.*, 1993).

Pe baza acestor parametrii s-a definit ***eroarea standard a mediei***, formula de calcul fiind:

$$e = \frac{\sigma}{\sqrt{n}} \qquad (7.1)$$

unde **σ** reprezintă abaterea standard a variabilei ***x*** pentru populația totală, abatare care de cele mai multe ori rămâne necunoscută, fiind înlocuită în calcule cu **s**, abaterea standard a aceleiaşi variabile într-un eşantion oarecare.

Pe baza erorii standard a mediei şi considerând că valorile medii, obținute pe o mulțime de eşantioane consecutive extrase din aceeaşi populație, sunt distribuite tot după curba normală a lui Gauss, putem stabili, cu o probabilitate de 95% sau 99%, limitele între care se găseşte adevărata valoare **μ** a colectivității generale. Intervalul delimitat de aceste limite este chiar *intervalul de încredere* stabilit pentru cele două *praguri (niveluri) de semnificație*:

- [m - 1,96e; m + 1,96e], interval de încredere la pragul de p = .05;
- [m - 2,58e; m + 2,58e], interval de încredere la pragul de p = .01.

Vom spune că există riscul ca în 5%, respectiv 1%, din cazuri adevărata medie să cadă în afara intervalului ales.

7.2.2. Semnificația frecvenței (absolute sau relative)

Analog, calculăm eroarea standard a frecvenței:

$$e = \sqrt{\frac{p \times q}{n}} \qquad (7.2)$$

unde **p** reprezintă chiar frecvența (cu condiția ca mărimea eşantionului să fie n>100) iar **q** = 1-*f*.

Intervalul de încredere va fi:

- [*f* - 1,96e; *f* + 1,96e], la pragul de p = .05;
- [*f* - 2,58e; *f* + 2,58e], la pragul de p = .01.

7.3. TESTAREA IPOTEZELOR

Testarea ipotezelor – demers fundamental în activitatea de cercetare ştiințifică – „reprezintă, alături de estimarea parametrilor statistici, unul dintre *principalele aspecte ale inferenței statistice*". (Dyer, 1995, apud Sava, 2004a, p. 27)

Ipoteza ştiințifică este o predicție care are capacitatea de a fi operaționalizată şi testată pentru a oferi un răspuns problemei studiate.

Modul de formulare a ipotezei cercetării determină două categorii de ipoteze:

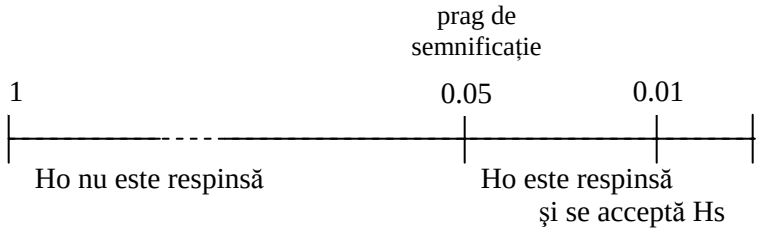
- unidirecționale (unilaterală), atunci când se precizează direcția predicției prin formulări de genul: „există o corelație pozitivă/negativă" sau „grupul A este mai bun/slab decât grupul B"

- bidirecționale (bilaterală), atunci când direcția predicției nu este precizată; vom avea formulări de genul: „există o corelație între variabile” sau „există diferențe între loturi”.

Dacă avem suficiente indicii cu privire la modul de evoluție a datelor este de preferat să optăm pentru formularea unor ipoteze unidirecționale, existând şanse mai mare ca aceasta să fie sprijinită. (Sava, 2004a)

Indiferent de modul de formulare, alături de această ***ipoteză specifică (Hs)***, - (numită şi ipoteză de cercetare, ipoteză de lucru sau ipoteză alternativă) se exprimă şi o altă ipoteză care să atribuie numai întâmplării, hazardului, tendințele sau diferențele constatate. Este vorba despre ***ipoteza nulă*** **(Ho)** (sau ipoteza statistică) asupra căreia se impun următoarele precizări:

- atât ipoteza nulă (Ho)cât şi ipoteza specifică (Hs) se referă la populație, nu la eşantioane ca atare;
- singurul lucru ce poate fi obținut prin testarea ipotezelor este respingerea sau nerespingerea ipotezei nule;
- dacă ipoteza nulă este respinsă, atunci ipoteza alternativă este sprijinită de datele obținute, altfel spus: ipoteza specifică este acceptată;
- decizia de a respinge ipoteza nulă se ia pe baza unui prag de semnificație (cel mai adesea .05 sau .01).



7.4. TESTELE PARAMETRICE t ŞI z.

Pe lângă studiul asocierii dintre variabile, tehnicile statistice pot fi utilizate şi pentru determinarea diferențelor dintre grupuri. Aceste metode se utilizează frecvent în cercetările experimentale.

Acest capitol prezintă acele tehnici parametrice care permit evaluarea efectelor unei variabile independente (manipulate de cercetător) sau categoriale (vârsta, sex, etc) asupra unei variabile dependente, în situația în care se lucrează cu una sau doua grupe de subiecți (Sava, 2004b).

Cu ajutorul acestor teste statistice se ridică problema dacă diferențele constatate între grupele de subiecți sunt datorate intervenției cercetătorului (variabilei independente), caracteristicilor variabilei categoriale sau dimpotrivă, întâmplării.

Există trei tipuri de tehnici principale:

1. Tehnici care privesc diferența dintre un eşantion şi media populației din care acesta face parte – „the one simple t Test”;
2. Tehnici care privesc diferența dintre două grupe independente de subiecți – „the t test for independent samples”;

3. Tehnici care privesc diferența dintre două grupe dependente de subiecți – „the t test for correlated samples”.

7.4.1. TEHNICILE t ŞI z PENTRU UN EŞANTION.

În acest caz dorim să aflăm dacă un eşantion de subiecți diferă de o populație mai mare. Să presupunem că un test de empatie a fost administrat pe o populație mare de subiecți elevi abia intrați la liceu (N = 1000), iar media obținută pe întreaga populație testată a fost de 76 (μ). Când s-a efectuat acelaşi test pe o clasă de elevi de n=32 subiecți, s-a obținut media de 81 (m) şi o estimare a abaterii standard de 9 (s). Se pune problema dacă elevii din această clasă au un nivel de empatie diferit de media specifică pentru clasa a IX-a.

Pentru soluționarea acestei probleme există două teste statistice adecvate, şi anume testele z şi t.

Vom utiliza testul z dacă:

- se cunoaşte abaterea standard a variabilei dependente la nivelul populației;
- numărul de subiecți cuprinşi în eşantionul comparativ este suficient de mare (de regulă peste 30 de subiecți).

În situația în care una din cele doua condiții nu este îndeplinită, utilizăm testul t (Student) pentru un eşantion.

În problema de față se observă că nu putem aplica testul z deşi avem un eşantion comparativ destul de mare n=32 (mai mare decât 30) deoarece nu se cunoaşte abaterea standard a populației din care face parte eşantionul.

Ca urmare, calculăm testul t care validează sau infirmă ipoteza nulă potrivit căreia, nu există nici o diferență între media (m) obținută pe eşantionul de subiecți (n= 32) şi media (μ) obținută pe populația din care a fost extras eşantionul.

Matematic, ipoteza nulă şi cea de lucru (alternativă) se formulează astfel:

Ho: $\mu = m$

Hs_1: $m \neq \mu$

Hs_2: $\mu > m$ ori $\mu < m$

În cazul Hs_1 ipoteza alternativă precizează existența unei diferențe între cele două medii fără a arăta direcția acestei diferențe. În acest caz avem de a face cu un test t bilateral (two-tailed test). În cazul Hs_2 ipoteza alternativă specifică direcția diferenței între cele două medii - o medie este mai mică (mare) decât cealaltă datorită unor considerente teoretice. Această situație necesită un test t unilateral (one-tailed).

Cele doua tipuri de test t utilizează aceeaşi formulă, specificul unilateral vs. bilateral influențând doar valorile comparative prezente în tabelul lui t (anexa 2).

Formula lui t este:

$$t = \frac{m - \mu}{EE_m} \qquad (7.3)$$

unde: m este media eşantionului

μ (miu) este media populației din care face parte eşantionul;

EE_m este eroarea standard a mediei eşantionului;

$$EE_m = \frac{s}{\sqrt{n}} \qquad (7.4)$$

unde: s este estimarea abaterii standard a eşantionului (s=9);
n este volumul (mărimea) eşantionului (n=32).

Calcularea testului z necesită utilizarea formulei:

$$z = \frac{m - \mu}{EE_\mu} \qquad (7.6)$$

unde: m este media eşantionului comparat;
μ este media populației;
EEμ este eroarea standard a mediei populației.

$$EE_\mu = \frac{\sigma}{\sqrt{n}} \qquad (7.7)$$

unde: σ (sigma) este abaterea standard a populației;
n este volumul eşantionului comparat.

Interpretarea valorii lui z obținute se face raportând această valoare la valorile standardizate ale lui z. Spre deosebire de testul t, care necesită consultarea tabelului t în vederea admiterii sau respingerii ipotezei nule, în cazul testului z, valoarea obținută se confruntă cu patru valori standardizate:

Testul bilateral: z = 1,96 pentru un $p < .05$
z = 2,58 pentru un $p < .01$
Testul unilateral: z = 1,65 pentru un $p < .05$
z = 2,33 pentru un $p < .01$

7.4.2. TESTELE t ŞI z PENTRU EŞANTIOANE INDEPENDENTE

Testele t şi z prezentate anterior pentru a determina dacă un eşantion diferă de o populație nu se aplică prea frecvent. Mai des sunt utilizate testele t şi z pentru a determina dacă mediile a două eşantioane, independente sau corelate (dependente), diferă semnificativ. Situațiile în care avem eşantioane independente sau dependente le-am prezentat în subcapitolul 7.1.

Ne punem întrebarea: „Când aplicăm testul t şi când aplicăm testul z?" Răspunsul ține de aceleaşi două condiții prezentate anterior: cunoaşterea abaterii standard a celor două eşantioane şi volumul acestora. Prima condiție este atinsă mult mai uşor, de aceea criteriul hotărâtor în alegerea tipului de test (t sau z) este volumul eşantionului. Există conform teoremei limitei centrale o evoluție a distribuției datelor în funcție de numărul de subiecți. Se consideră şi se acceptă de majoritatea cercetătorilor, că un eşantion de 30 de subiecți sau mai mult are o distribuție normală a datelor z. Un număr mai mic de 30 de subiecți determină o distribuție asimetrică a datelor de tip t. Chiar dacă se utilizează o împărțire grosieră, s-a stabilit de către cercetători următoarea clauză pentru cazul a două eşantioane:

- Dacă $n_1 < 30$ (numărul de subiecți din prima grupă) şi $n_2 < 30$ (numărul de subiecți din a doua grupă) se aplică testul t.
- Dacă $n_1 > 30$ şi $n_2 > 30$ se aplica testul z.

TESTUL t (STUDENT) INDEPENDENT

Testul t independent.

$$t = \frac{m_I - m_{II}}{EE_{m_I - m_{II}}} \tag{7.8}$$

unde: m_I şi m_{II} reprezintă mediile celor două eşantioane;
$EE_{m_I-m_{II}}$ reprezintă eroarea standard a diferenței dintre cele două medii.

Pentru calculul erorii standard a diferenței dintre medii ($EE_{m_I-m_{II}}$) folosim formulele:

Dacă n_I este egal n_{II}:

$$EE_{m_I - m_{II}} = \sqrt{\frac{s_I^2}{n_I} + \frac{s_{II}^2}{n_{II}}} \tag{7.9}$$

$$EE_{m_I - m_{II}} = \sqrt{\frac{\sum x_I^2 - \frac{(\sum x_I)^2}{n_I} + \sum x_{II}^2 - \frac{(\sum x_{II})^2}{n_{II}}}{n_I (n_{II} - 1)}} \tag{7.9'}$$

unde: s_I^2 reprezintă dispersia primului grup (abaterea standard la pătrat); s_{II}^2 reprezintă dispersia celui de-al doilea grup; n_I - numărul de subiecți din primul grup; n_{II} - numărul de subiecți din al doilea grup.

Dacă n_I este diferit de n_{II}:

$$EE_{m_I - m_{II}} = \sqrt{\left(\frac{\sum x_I^2 - \frac{(\sum x_I)^2}{n_I} + \sum x_{II}^2 - \frac{(\sum x_{II})^2}{n_{II}}}{n_I + n_{II} - 2} \right) \left(\frac{1}{n_I} + \frac{1}{n_{II}} \right)} \tag{7.10}$$

TESTUL Z INDEPENDENT

În situația în care $n_I > 30$ şi $n_{II} > 30$ şi a două eşantioane independente aplicăm testul z. Formula de calcul este:

$$z = \frac{m_I - m_{II}}{\sqrt{\frac{s_I^2}{n_I} + \frac{s_{II}^2}{n_{II}}}} \tag{7.12}$$

După cum se observă formula de calcul a lui z în această situație este identică cu cea a lui t independent pentru $n_I = n_{II}$. Spre deosebire de testul t independent, testul z are aceeaşi formulă şi în cazul în care $n_I \neq n_{II}$.

Rezultatul obținut este comparat cu cele două valori standardizate z (1,96 pentru $p < .05$, respectiv 2,58 pentru $p < .01$ pentru testul bilateral, respectiv cu 1,65 pentru $p < .05$, respectiv 2,33 pentru $p < .01$ pentru testul unilateral). Algoritmul rezolvării problemelor care necesită testul z este asemănător cu cel prezentat în cazul lui z pentru un eşantion.

7.4.3. TESTELE t ŞI z PENTRU EŞANTIOANE DEPENDENTE

Se folosesc atunci când elementele componente ale celor două grupe sunt în relație de corespondență.

Formula lui t dependent este:

$$t = \frac{m_I - m_{II}}{EE_d} \qquad (7.13)$$

unde: m_I şi m_{II} sunt mediile celor două grupe;
EE_d este eroarea standard a diferenței (d).

Pentru a calcula EE_d utilizăm una din formulele:

$$EE_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} \qquad (7.14)$$

unde: d este diferența dintre pre-test şi post-test, între poziția unu în prima grupă şi poziția unu din a doua grupă ş.a.m.d.;
n este numărul de perechi de subiecți (în cazul problemei date 12).

sau

$$EE_d = \sqrt{\frac{s_I^2}{n_I} + \frac{s_{II}^2}{n_{II}} - 2r_{12} * \frac{s_I}{n_I} * \frac{s_{II}}{n_{II}}} \qquad (7.14')$$

unde: s_I^2 şi s_{II}^2 sunt dispersiile celor două grupe;
n_I şi n_{II} sunt egale şi reprezintă numărul de perechi de subiecți;
r_{12} este coeficientul de corelație între datele celor două grupe;
s_I şi s_{II} sunt abaterile standard ale celor două grupe.

TESTUL z DEPENDENT

Acesta poate fi utilizat în cazul eşantioanelor mai mari de 30 de subiecți fiecare. În această situație EE_d (eroarea standard a diferenței) se calculează utilizând formula 7.14' prezentată pentru t dependent care conține coeficientul de corelație r_{12}.

Interpretarea rezultatului obținut se face după acelaşi algoritm prezentat şi la celelalte teste z pentru un eşantion şi două eşantioane independente.

Considerațiile făcute în cazul testului z independent cu privire la tendința actuală de a înlocui testul z cu testul t chiar în cazul eşantioanelor mai mari de 30 de subiecți rămâne validă şi pentru testele dependente.

7.5. UTILIZAREA SPSS PENTRU APLICAREA TESTULUI t

1. TESTUL t PENTRU MEDIA UNUI SINGUR EŞANTION

Se parcurge, în bara de meniuri, traseul:

„Analyze" – „Compare Means" – „One-Sample T Test..."

Va fi afişată fereastră de dialog intitulată *„One-Sample T Test"* (figura 7.1).

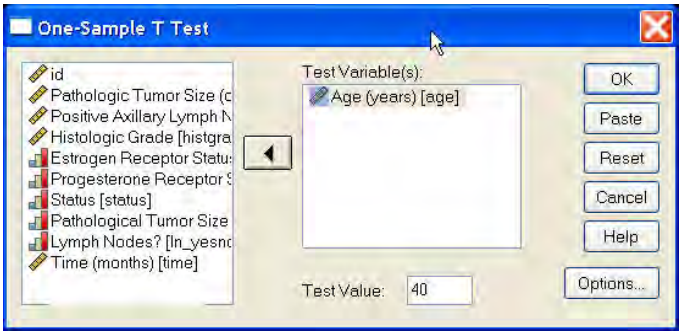
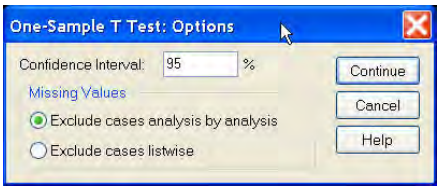


Figura 7.1. *Fereastra pentru calculul testului t pentru media unui singur eşantion.*

Vom începe prin a selecta variabila testată mutând-o din partea stângă în fereastra *„Test Variable(s)"*. În zona *„Test Value"* se înscrie media populației, sau altă valoare de referință.

Prin apăsarea butonului *„Options"* se va deschide o nouă fereastră în care vom putea schimba valoarea pragului de semnificație. Confidence Interval 95% este echivalent cu p=0.05 si este valoarea implicită pentru toate testele statistice.



Apăsăm *„Continue"* iar în final *„OK"*.

*

2. TESTUL t PENTRU EŞANTIOANE INDEPENDENTE

Se parcurge, în bara de meniuri, traseul:

„Analyze" – „Compare Means" – „One-Sample T Test..."

Va fi afişată fereastră de dialog intitulată *„One-Sample T Test"* (figura 7.1).

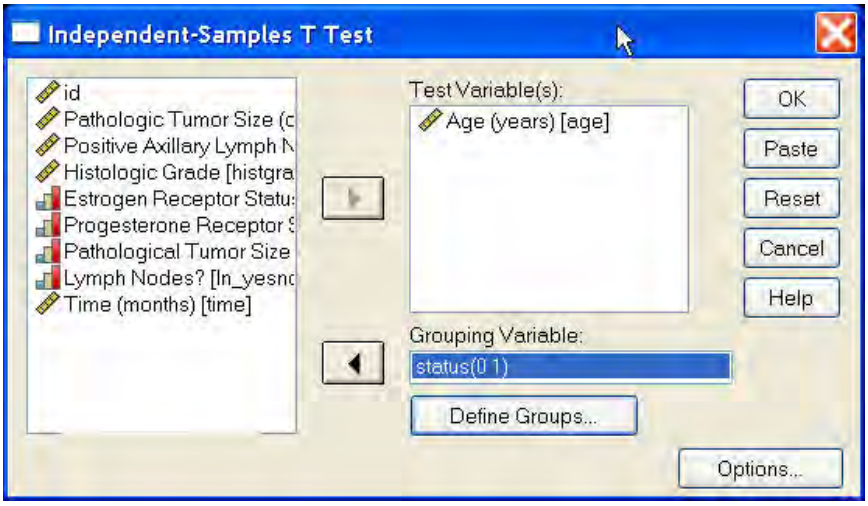
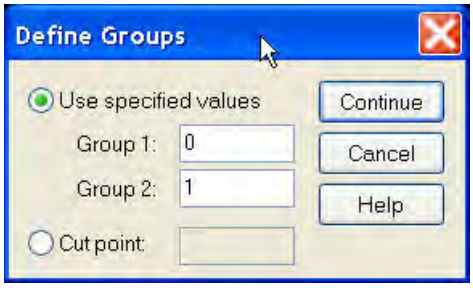


Figura 7.2. *Fereastra pentru calculul testului t pentru eşantioane independente.*

Şi de data această vom începe prin a selecta variabila testată mutând-o din partea stângă în fereastra „*Test Variable(s)*”. Diferența apare în zona „*Grouping Variable*”, acolo unde va trebui să definim variabila independentă (grup), cea care face diferența între eşantioanele independente.



Prin apăsarea butonului „*Define Groups*” se va deschide o nouă fereastră în care vom specifica valorile care definesc cele două grupuri.

Apăsăm „*Continue*”, iar dacă toate câmpurile le-am completat corect se va activa butonul „*OK*”.

*

3. TESTUL T PENTRU DIFERENȚA DINTRE MEDIILE A DOUĂ EŞANTIOANE DEPENDENTE (PERECHI)

Se parcurge, în bara de meniuri, traseul:

„*Analyze*” – „*Compare Means*” – „*Paired-Sample T Test...*”

Va fi afişată fereastră de dialog intitulată „*Paired -Sample T Test*” (figura 7.3).

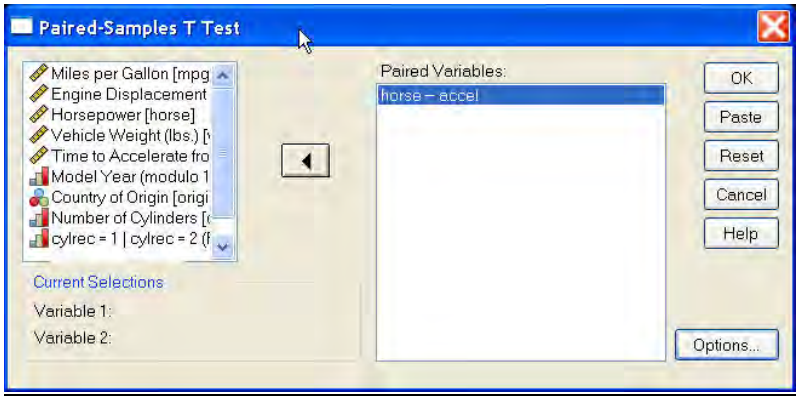
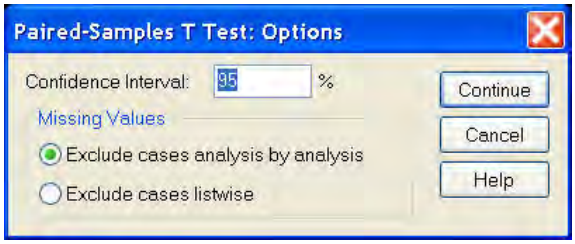


Figura 7.3. *Fereastra pentru calculul testului t pentru eşantioane perechi.*

Se selectează cu câte un clic de mouse, pe rând, fiecare dintre cele două variabile. Astfel se constituie perechea de variabile în zona „*Current selection*”. O dată constituită, perechea de variabile se trece în lista „*Paired Variables*” cu butonul de transfer (►). Pot fi create mai multe perechi de variabile şi prelucrate simultan.



Caseta „*Options*” permite alegerea pragului de semnificație, dacă dorim schimbarea celui implicit (p=0.05).

8.

CORELAȚIE ŞI REGRESIE

8.1. Noțiunea de covarianță
8.2. Coeficienții de corelație
  8.2.1. Clasificarea coeficienților de corelație
  8.2.2. Formula coeficientului de corelație liniară simplă (Bravais-Pearson)
  8.2.3. Reprezentarea grafică a corelației. Liniaritatea relației.
  8.2.4. Interpretarea coeficientului de corelație. Mărimea efectului.
8.3. Coeficienți de corelație parametrici
  8.3.1. Coeficientul de corelație Pearson r
  8.3.2. Coeficientul r_{bis}
8.4. Coeficienți de corelație neparametrici:
  8.4.1. Coeficientul de corelație a rangurilor Spearman ρ
8.5. Regresia simplă liniară.
8.6. Utilizarea SPSS pentru determinarea coeficienților de corelație

Adesea, în practica sau cercetarea psihologică, pe lângă aplicarea testelor de semnificație prezentate în capitolul anterior (prin care verificăm semnificația diferenței între două medii ale aceleiaşi variabile, măsurate în două situații diferite), suntem interesați de *gradul de asociere dintre două variabile* măsurate pe acelaşi grup de subiecți. De data aceasta vom opera cu mai mult de o singură variabilă. Vorbim astfel de o *statistică bivariată*, axată pe indicatori descriptivi de asociere sau de relaționare, înțeleşi prin termenii de covarianță şi independență.

8.1. NOȚIUNEA DE COVARIANȚĂ

Covarianța este rezultatul variației concomitente a valorilor care aparțin de două variabile. Covarianța ne indică existența unei legături între variația valorilor unei variabile în raport cu cealaltă variabilă. De exemplu (adaptare după Radu *et.al.*, 1993, p.103), observând notele obținute de aceiaşi elevi la matematică şi la fizică, constatăm că ele covariază, adică sunt asemănătoare: elevii cu performanțe notabile la matematică au note mari şi la fizică, şi reciproc. În realitate, situațiile de acest gen sunt foarte multe: nivelul ridicat al pregătirii şcolare covariază cu numărul de cărți citite într-o perioadă de timp; performanțele ridicate în conducerea autovehiculului sunt asociate cu rezultatele ridicate la testele de atenție etc.

Conceptul de **independență** se opune celui de covarianță. El este caracteristic unei situații de neasociere între două variabile. Independența se referă la relația dintre două evenimente, variabile sau seturi de date, astfel încât nici una nu poate fi

influențată de alta şi schimbările care pot fi realizate la nivelul uneia sunt posibile fără să o influențeze pe cealaltă (English & English, 1958, apud Pitariu, 1991). Desigur, independența trebuie luată în sens relativ. De exemplu, nu putem considera ca asociere relația dintre inteligență şi numărul copacilor dintr-o pădure.

Covariația dintre două variabile poate fi evidențiată prin trei elemente descriptive (Sava, 2004):

- calcularea coeficienților de corelație,
- reprezentarea grafică a norului de puncte,
- realizarea de tabele de contingență (de asociere).

8.2. COEFICIENȚII DE CORELAȚIE

Coeficienții de corelație sunt indicatori descriptivi ce arată gradul de covariație dintre două variabile. Ei reflectă gradul de variație concomitentă dintre două şi numai două variabile: o singură variabilă independentă (*X*) şi o singură variabilă dependentă (*Y*). Când cele două variabile covariază în acelaşi sens, vorbim despre o corelație *pozitivă* (ex. cu cât timpul alocat pregătirii examenului de statistică este mai mare, cu atât nota obținută la evaluarea finală este mai bună). Dacă asocierea este în direcții opuse (în timp ce o variabilă creşte, cealaltă scade), discutăm despre o corelație *negativă*. (ex. performanța unui angajat la un test de atenție concentrată este cu atât mai bună cu cât numărul de erori este mai mic).

Se impune o precizare. Spre deosebire de experiment, care dezvăluie relații cauză-efect, studiul de corelație nu oferă nemijlocit o măsură a cauzalității, ci pur şi simplu a modului de asociere. Coeficientul de corelație este un index al prezenței/absenței unei relații între două variabile şi nu un index al unei relații cauzale. Corelația însă este implicată în predicție. O corelație semnificativă (mare) între *X* şi *Y* ne poate spune, cu diferite grade de precizie că prin cunoaşterea valorii uneia dintre cele două variabile, putem să estimăm valoarea celeilalte (ex. dacă scorurile la unele scale din CPI (*Y*) sunt ridicate, atunci şi performanțele manageriale (*X*) se poate estima că vor fi ridicate; condiția este ca între cele două variabile să existe o corelație semnificativă.)

8.2.1. Clasificarea coeficienților de corelație

Coeficienții de corelație se împart în două mari categorii:

- coeficienți de corelație parametrici: coeficientul Bravais-Pearson (r), biserial (r_{bis}), punct biserial (r_{pbis});
- coeficienți de corelație neparametrici: coeficientul de corelație a rangurilor Spearman (ρ), coeficientul Kendall (τ), .

În funcție de tipul datelor colectate şi de liniaritatea/monotonia relației dintre cele două variabile, tratatele de statistică prezintă o multitudine de coeficienți de corelație. Ne vom limita în această lucrare doar la prezentarea celor care sunt utilizați mai des de către psihologi şi pedagogi.

Tabelul 8.1 Utilizarea coeficienților de corelație în funcție de tipul variabilelor[1].

		Variabila independentă *x*			
		Nominală **dihotomică**	**Nominală** cu mai mult de două valori	**Ordinală**	**Numerică** (de interval sau de raport)
Variabila dependentă *y*	Nominală **dihotomică**	r, φ, χ^2, $r_{tetrahoric}$	χ^2, λ, C, V	Kendall τ	r, r_{bis}, r_{pbis}
	Nominală cu mai mult de două valori		χ^2, λ, C, V	Chi pătrat χ^2, λ	χ^2, λ
	Ordinală			Spearman ρ Kendall τ	Spearman ρ Kendall τ
	Numerică (de interval sau de raport)				Person r

8.2.2. Formula de calcul a coeficientului de corelație liniară simplă

După cum ştim, coeficienții de corelație ne arată gradul de covariație dintre două serii statistice. Covarianța dintre variabila *X* şi variabila *Y* ne este dată de formula:

$$\text{cov}_{xy} = \frac{\sum x \cdot y}{n} \qquad (8.1)$$

În această formulă, *x* şi *y* sunt valorile-pereche ale celor două variabile, iar *n* reprezintă volumul eşantionului. Deşi reflectă cu succes asocierea sau relaționarea dintre cele două variabile, calculul covarianței întâmpină o problemă: produsul de la numărător are sens doar dacă cele două variabile sunt exprimate în aceeaşi unitate de măsură. De exemplu (Popa, 2009), este evident faptul că, nu putem aplica formula de mai sus pentru a studia covarianța dintre înălțime şi greutate, deoarece este dificil să înțelegem rezultatul unui produs dintre unități de măsură diferite (kg pentru greutate şi cm pentru lungime). Acest inconvenient a fost eliminat prin transformarea valorilor celor două variabile în cote z. Astfel, produsul scorurilor standard z_x şi z_y nu mai are legătură cu unitățile de măsură ale lui *X* şi *Y*. Mai mult, această standardizare (i) va egaliza influența variabilelor asupra gradului de asociere dintre ele (de exemplu [Sava, 2004], dacă vom calcula covarianța dintre venit şi numărul anilor de şcoală absolviți, prima variabilă, având o amplitudine mai mare, va contribui mai mult la rezultatul final; venitul poate varia între 0 şi 10.000, în timp ce numărul anilor de şcoală absolviți poate fi de maxim 25) şi (ii) va permite compararea gradului de asociere dintre două variabile cu asocierea dintre alte două variabile (de exemplu, care asociere este mai puternică, între inteligența băieților şi a taților sau între frumusețea fetelor şi a mamelor?!).

În consecință, corelația este o formă standardizată a covarianței, eliminând problema măsurării datelor prin scale diferite. Formula de calcul a corelației este:

$$r = \frac{\sum z_x \cdot z_y}{n} \qquad (8.2)$$

[1] Literele greceşti din tabel au următoarele pronunții: χ^2=chi pătrat, ρ=rho, τ=tau, λ=lamda, φ=phi.

unde z_x şi z_y scorurile z ale variabilelor *X* şi *Y*, iar *n* mărimea eşantionului.

r exprimă intensitatea relației liniare dintre valorile a două variabile şi este cunoscut sub numele de ***coeficient de corelație liniară simplă***. Îl mai găsim sub denumirile: coeficient de corelație al „moment-produsului", coeficient de corelație Bravais-Pearson[2] sau chiar simplu „Pearson r".

Coeficientul de corelație Bravais-Pearson are cea mai mare frecvență de utilizare în psihologie, însă -atenție!- se foloseşte doar când relația dintre variabilele supuse calculului de corelație este liniară (vezi 8.2.3.), iar cele două variabile sunt exprimate numeric (în puține cazuri, acceptăm şi variabile măsurate prin scale nominale dihotomice).

Valorile lui *r* sunt cuprinse între -1 şi +1, trecând prin 0 care indică absența corelației. Dacă *r* este pozitiv, atunci vorbim de o corelație directă, pozitivă. În cazul acesta, dacă una din variabile *X* creşte, atunci şi cealaltă variabilă *Y* va avea tendința de a creşte.

Când coeficientul de corelație este nul, se spune doar că variabilele *X* şi *Y* sunt necorelate, eventual independente.

Dacă *r* este negativ, atunci *Y* va avea tendința de a varia în medie sens invers lui *X*. În acest caz corelație este negativă, inversă.

Valorile *r* = -1 şi *r* = +1 ne indică existența unei relații perfecte între variabile.

-1	0	+1
Asociere negativă (inversă)	Lipsă de asociere	Asociere pozitivă (directă)

Figura 8.1. *Valorile coeficienților de corelație*

Formula coeficientului de corelației ia în considerare, de fiecare dată, câte două variabile statistice. De multe ori, în studiile psihosociale ne interesează asocierea dintre mai multe variabile. Spre exemplu, dacă avem trei variabile X, Z, şi Z vom calcula succesiv r_{xy}, r_{xz} şi r_{yz}. Cu aceste valori putem întocmi o matrice a coeficienților de corelație utilizată în analiza factorială.

8.2.3. Reprezentarea grafică a corelației. Liniaritatea relației.

În cercetarea psihologică a corelației, analiza **norului de puncte**[3] este de mare importanță, oferind numeroase explicații suplimentare față de un simplu coeficient de corelație. Astfel, ni se oferă detalii referitor la forma relației dintre două variabile (liniară sau neliniară – figura 8.2.), direcția (pozitivă, negativă sau absența unei asocieri– figura 8.3.), intensitatea relației dintre două variabile (puternică, medie sau

[2] La sfârşitul secolului al XIX-lea, statisticianul englez Karl Pearson (1857-1936) dezvoltă, prin utilizarea datelor cuprinse în încercările lui Bravais, forma finală a coeficientului de corelație prin momentul produselor. Pearson fost elev al celebrului matematician Francis Galton (1822-1911), cel care a introdus tehnica corelației în biologie şi psihologie. (Clocotici & Stan, 2001)

[3] În engleză *scatterplot.*

scăzută). O incursiune în domeniul reprezentării grafice a coeficientului de corelație o găsim deci utilă.

Examinarea norului de puncte, care reprezintă proiecția fiecărui subiect într-un spațiu bidimensional, se poate afirma că este un pas semnificativ în studiul corelației dintre două variabile. El oferă, în final, indicii asupra tipului de coeficient de corelație pe care dorim să-l calculăm.

8.2.4. Interpretarea coeficientului de corelație. Mărimea efectului.

Interpretarea încrederii lui r

Criteriul după care poate fi discutată semnificația lui r presupune consultarea unei tabele special construite. Prin acest procedeu se poate respinge ipoteza nulă conform căreia nu există o relație adevarată (semnificativă), între variabile, iar eventualele asocieri se datorează întâmplării. Dacă o relație este semnificativă din punct de vedere statistic, adică este de încredere, înseamnă ca vom obține rezultate similare dacă s-ar reface experimentul.

În utilizarea tabelului lui r putem alege diferite praguri de semnificație. Există o înțelegere la nivelul comunității ştiințifice internaționale cum că pragul minim acceptat pentru a considera o relație semnificativă statistic este 0,05. Aceste valori pot fi însa şi mai mici.

Pentru aflarea semnificației unui coeficient de corelație este necesară parcurgerea următorilor paşi:

1. Se alege nivelul de semnificație dorit, să zicem de 0,05.
2. Se stabileşte tipul de relație între variabile: bilaterală (two-tailed), respectiv unilaterală (one-tailed).
3. Se citeşte din tabel (Anexa 3) valoarea lui r pentru coloana corespunzătoare numărului de grade de libertate (notat cu df). Acestea sunt pentru r de df=N-2 stabilindu-se în funcție de numărul de subiecți N validați.
4. Daca valoarea lui r obținută în urma calculării sale o depăşeşte pe cea din tabel, atunci aceasta este semnificativă la pragul de semnificație ales, în cazul nostru de 0,05 (notat şi cu .05) şi numărul de grade de libertate specificat.

Interpretarea corelației din perspectiva semnificației

Statistica poate răspunde la două întrebări privind datele pe care le avem: Sunt autentice relațiile (efectele) descoperite? Ce semnificație au acestea?

Cel mai utilizat criteriu pentru interpretarea semnificației coeficientului de corelație este coeficientul de determinare (r^2 – r pătrat). Acest criteriu nu are întotdeauna însemnătate din cauza influenței importante pe care o are mărimea lotului în determinarea coeficientului de corelație. El trebuie analizat cu grija în cazurile în care exista un număr relativ mic de subiecți (sub 20). De asemenea, coeficientul de determinare poate fi aplicat doar dacă am obținut în prealabil un r semnificativ.

Prin intermediul lui r pătrat se determina partea de asociere comună a factorilor care influențează cele doua variabile. Cu alte cuvinte, coeficientul de determinare

indică partea din dispersia totală a măsurării unei variabile care poate fi explicată sau justificată de dispersia valorilor din cealaltă variabilă.

De exemplu, dacă într-un studiu corelația găsită a fost de 0,83, atunci putem afirma că $r^2 = (r)^2$ (coeficientul de corelație la pătrat) este de 0,69. Uzual coeficientul de determinare se înmulțește cu 100 şi exprimarea se transforma în procente din dispersie (69%).

8.3. COEFICIENȚI DE CORELAȚIE PARAMETRICI

Pentru a calcula coeficienții de corelație parametrici, variabilele studiate trebuie să îndeplinească următoarele condiții:

- să fie variabile numerice (exprimate pe scale de intervale sau de rapoarte),
- variabila supusă studiului să aibă o distribuție cât mai apropiată de cea normală și un grad ridicat de omogenitate;
- distribuția comună a variabilelor să nu prezinte valori extreme (outliers).

Verificarea acestor condiții este o etapă preliminară în orice analiză bazată pe studiul corelațional. Este important de reținut că, înainte de a calcula unul sau altul dintre coeficienți, trebuie să verificăm valorile mediei, abaterii standard şi a indicatorilor de asimetrie, să analizăm norul de puncte ce reprezintă grafic asocierea dintre variabile, iar, dacă este cazul, să eliminăm valorile extreme[4] sau să asigurăm condiția de homoscedasticitate[5].

8.3.1. Coeficientul de corelație Pearson r.

Atunci când variabilele sunt prezentate sub formă de scoruri brute, formula de calcul a lui Pearson r, este următoarea:

$$r = \frac{\Sigma(x_i - m_x)(y_i - m_y)}{\sqrt{\Sigma(x_i - m_x)^2 \cdot \Sigma(y_i - m_y)^2}} \qquad (8.3)$$

Aceasta este o formulă derivată din (8.2), în care s-au înlocuit expresiile pentru scorurile z_x şi z_y . Putem să simplificăm calculele utilizând o formulă asemănătoare, care se bazează pe calcule mai uşor de realizat:

$$r = \frac{\Sigma(x_i - m_x)(y_i - m_y)}{n \cdot s_x \cdot s_y} \qquad (8.4)$$

8.3.2. Coeficientul r biserial

Coeficientul r biserial îl găsim notat cu simbolul r_b sau r_{bis}. Este utilizat când două variabile corelabile sunt continue, dar una din ele a fost arbitrar dihotomizată. Există exemple numeroase când într-o cercetare corelațională este mai avantajos să

[4] Le mai putem spune valori neobişnuite sau influente; în engleză se numesc „*outliers*".

[5] Este o proprietate a relației liniare dintre două variabile exprimată prin omogenitatea norului de puncte ce reprezintă distribuția comună a variabilelor.

împărțim distribuția scorurilor în două clase, nu neapărat egale. Uneori chiar suntem constrânşi de împrejurări să facem acest lucru, neavând la dispoziție decât o singură variabilă, cum ar fi de pildă situația de „acceptat”/„respins” la un test de cunoştințe profesionale; această dihotomie o mai putem realiza în funcție de comportamentul „extravertit”/„intravertit”, de *locusul controlului* „intern”/„extern” etc.

Formula coeficientului *r* biserial, utilizat când avem de-a face cu variabile dihotomice sau organizate pe mai multe clase, este următoarea:

$$r_{bis} = \frac{m_p - m_q}{\sigma_t} \times \frac{pq}{y} \qquad (8.5)$$

unde: m_p = media scorurilor celor declarați „acceptați” la testul profesional;
m_q = media grupului celor „respinşi” la testul profesional;
p = proporția în grupul celor „acceptați”; *q* = (1-*p*) proporția celor „respinşi”
σ_t = abaterea standard pe lotul total;
y = ordonata unității de arie a curbei normale la punctul care împarte aria totală în două segmente (p+q=1) – valoarea *pq/y* se extrage din tabele.

OBSERVAȚIE: În cazul coeficientului de corelație biserial numărul de subiecți cuprinşi în eşantion trebuie să fie mai mare de 50.

8.4. COEFICIENȚI DE CORELAȚIE NEPARAMETRICI

Coeficientul de corelație Bravais-Pearson nu poate fi utilizat în orice situație. Apelul în orice condiții la acesta este o eroare pe care o fac mulți psihologi când vor să facă un studiu corelațional. Un criteriu important în alegerea metodei adoptate în calculul coeficientului de corelație este analiza atentă a setului de date cu care se operează. În continuare vom menționa câteva situații particulare în care sunt folosiți alți coeficienți de corelație decât *r*.

8.4.1. Coeficientul de corelație a rangurilor rho sau ρ (Spearman)

Când o scală (ex. variabila *X*) este o măsură ordinală şi când a doua scală (ex. *Y*) este fie o scală ordinală, fie una de raport sau de interval, nu se poate calcula coeficientul de corelație *r* a lui Bravais-Pearson.

Coeficientul de corelație ρ se bazează pe calculul diferenței de ranguri obținute de subiecți la cele două variabile. Formula de calcul este următoarea:

$$\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)} \qquad (8.7)$$

unde D reprezintă diferența de rang obținută pe cele două variabile, pentru fiecare observație în parte.

Coeficientul de corelație a rangurilor Spearman ρ are acelaşi domeniu de variație (-1/+1) şi se interpretează în acelaşi mod ca şi coeficientul de corelație pentru date parametrice Pearson *r*.

8.5. REGRESIA SIMPLĂ LINIARĂ

Într-un sens larg, *regresia* este o analiză a relației existente între variabile. O ecuație de regresie simplă conține o variabilă independentă (X) şi o variabilă dependentă (Y). O ecuație care conține mai multe variabile independente este o *ecuație de regresie multiplă* (R). Dacă procedăm la reprezentarea grafică a corelației dintre două variabile distribuite liniar, observăm că norul de puncte poate fi divizat de o dreaptă, *linia de regresie* sau „linia celei mai bune predicții". Prin intermediul acestei linii, pot fi făcute predicții asupra cărei valori a lui X îi va corespunde o valoare a lui Y (şi invers). Utilitatea practică cea mai importantă a folosirii ecuației de regresie în testarea psihologică, este să facă o predicție a unui scor sau altă variabilă, când este cunoscută o variabilă. Cu cât corelația dintre două variabile este mai mare, cu atât predicția va fi mai precisă. (Pitariu, 1991)

Formula ecuației de predicție este:

$$Y = a + bX \qquad (8.8)$$

În formula de mai sus, *a* şi *b* sunt *coeficienții de regresie*; *b* se referă la panta liniei de regresie iar *a* este o constantă. Ambii coeficienți se pot determina pe baza unor calcule algebrice din datele brute.

8.6. UTILIZAREA SPSS PENTRU CALCULAREA COEFICIENȚILOR DE CORELAȚIE

Se parcurge, în bara de meniuri, traseul:

„Analyze" – „Corelate" – „Bivariate..."

Va fi afişată fereastră de dialog intitulată „*Bivariate Correlations*" (figura 8.5).

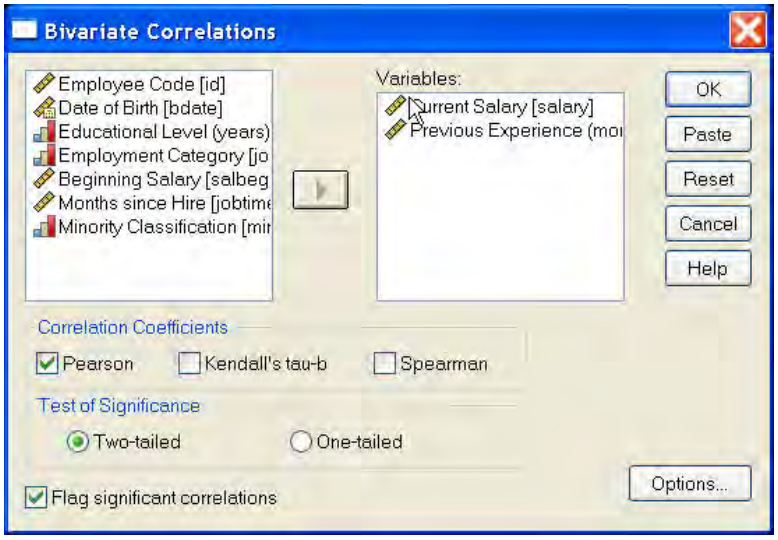


Figura 8.5. *Fereastra pentru calculul coeficienților de corelație.*

Vom începe prin a selecta variabilele supuse corelației mutându-le din partea stângă în fereastra „*Variables:*". Pot fi selectate mai mult de două variabile, situație în care vom obține coeficienții de corelație pentru toate perechile posibile de câte două variabile. De exemplu, dacă selectăm trei variabile X, Y şi Z, vom obține r_{xy}, r_{xz} şi r_{yz}.

În zona „*Correlation Coefficients*", în mod implicit va fi selectat coeficientul Pearson (r). Dacă variabilele nu sunt distribuite normal sau dacă sunt măsurate pe scale ordinale (neparametrice), vom selecta fie coeficientul de corelație a lui Kendal (τ), fie pe cel al lui Spearman (ρ).

La rubrica „*Test of Significance*", tipul implicit de testare a ipotezei este bilateral („*Two-tailed*"), dar se poate alege unilateral („*One-tailed*").

„*Flag significant correlations*", are ca efect marcarea cu un asterisc a coeficienților semnificativi la p=0.05 şi cu două asteriscuri a celor semnificativi la p=0.01. Acest lucru este util atunci când matricea de corelație este mare, pentru a scoate în evidență valorile semnificative ale lui r.

Apăsând butonul „*Options...*" putem solicita calcularea altor indicatori statistici ai variabilelor respective (de exemplu: media şi abaterea standard).

*

REPREZENTAREA GRAFICĂ A CORELAȚIEI CU AJUTORUL SPSS (SCATTERPLOT)

Pentru a vizualiza norul de puncte, implicit pentru a stabili caracterul şi intensitatea corelației dintre cele două variabile folosim o procedură grafică specifică, numită ***scatterplot***.

În bara de meniuri a programului SPSS vom parcurge traseul:

„Graphs" – „Legacy Dialogs" – „Scatter/Dot..."

Se va deschide o fereastră nouă din care selectăm „*Simple Scatter*".

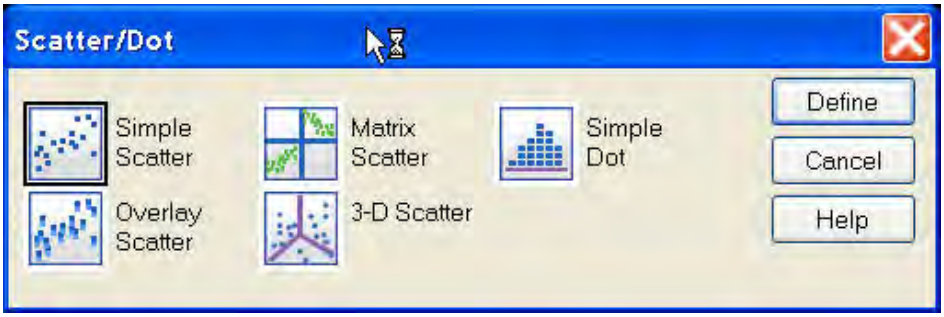


Figura 8.6. *Fereastra în care selectăm modalitatea de reprezentare scatterplot.*

BIBLIOGRAFIE

Bădiță, Maria şi Cristache, Silvia Elena (1998) – *Statistică – aplicații practice*. Bucureşti, Editura Mondan.

Biji, Mircea şi Biji, Elena (1979) - *Statistică teoretică*. Bucureşti, Editura Didactică şi Pedagogică.

Blezu, Dorin (2002) – *Statistica*. Sibiu, Editura Alma Mater.

Boudon, Raymond (1971) – *Les mathematiques en sociologie*. Paris, PUF.

Clocotici, Valentin şi Stan, Aurel (2000) – *Statistică aplicată în psihologie*. Iaşi, Polirom.

Cramer, Duncan (1994) – *Introducing Statistics for Social Research*. London, Routledge.

Culic, Irina (2004) – *Metode avansate în cercetarea socială Analiza multivariată de interdependență*. Iaşi, Polirom.

Dragoman, Dragoş (2003) – *Metode de analiză aplicate în ştiințele politice*. Sibiu, Continent.

Giulvezan, C., Zaporojan, G. şi Grindeanu, S. (2000) – *Introducere în informatica socială*. Timşoara, Editura de Vest.

Gravetter, F.J. şi Wallnau, L.B. (1992) – *Statistics for the Behavioral Sciences (3rd ed.)*. St. Paul, West Publishing Company.

Hartley, Alick (1999) – *Bazele statisticii*. Bucureşti, Editura Niculescu.

Jaba, Elisabeta şi Grama, Ana (2004) – *Analiza statistică cu SPSS sub Windows*. Iaşi, Polirom.

Luduşan, Nicolae şi Voiculescu, Florea (1997) - *Măsurarea şi analiza statistică în ştiințele educației*. Sibiu, Editura IMAGO.

Mărginean, Ioan (1982) – *Măsurarea în sociologie*. Bucureşti, Editura Ştiințifică şi Enciclopedică.

Novak, Andrei (1995) - *Statistică socială aplicată*. Bucureşti, Editura Hyperion.

Pitariu, Horia (1991) – *Introducere în statistica psihologică şi educațională*. Cluj-Napoca, Universitatea „Babeş-Bolyai" din Cluj-Napoca.

Popa, Marian (2009) – *Statistică pentru psihologie. Teorie şi aplicații SPSS*. Iaşi, Polirom.

Popa, Marian (2004) – *Statistică psihologică cu aplicații SPSS*. Bucureşti, Editura Universității din Bucureşti.

Popa, Marian (2006) – *Statistică psihologică – Curs de bază*. Găsită la http://popamarian.googlepages.com.

Popescu, Angela (2000) - *Statistică*. Bucureşti, Editura Fundației *România de Mâine*.

Porojan, Dumitru (1993) - *Statistica şi teoria sondajului*. Bucureşti, Casa de editură şi presă „Şansa" S.R.L..

Radu I. (coord.) (1993) – *Metodologia psihologică şi analiza datelor*, Cluj-Napoca, Editura Sincron.

Rateau, Patrick (2004) – *Metodele şi statisticele experimentale în ştiințele umane*. Iaşi, Polirom.

Rotariu, Traian (coord.) (1999) – *Metode statistice aplicate în ştiințele sociale*. Iaşi, Polirom.

Sandu, Dumitru (1992) – *Statistica în ştiințele sociale*, Universitatea din Bucureşti.

Sava, Florin (2004a) – *Analiza datelor în cercetarea psihologică*. Cluj-Napoca, Editura A.S.C.R.

Sava, Florin (2004b) – *Pagina de statistică socială*. Găsită la http://statisticasociala.tripod.com.

Simion, Doina Maria (2002) – *Bazele statisticii*. Sibiu, Editura Alma Mater.

* * * – *SPSS 7.5 for Windows - Brief Guide*. Chicago, Prentice-Hall Inc., 1997.

Yule, G.U. şi Kendall, M.G. (1969) – *Introducere în teoria statisticii*. Bucureşti, Editura Ştiințifică.